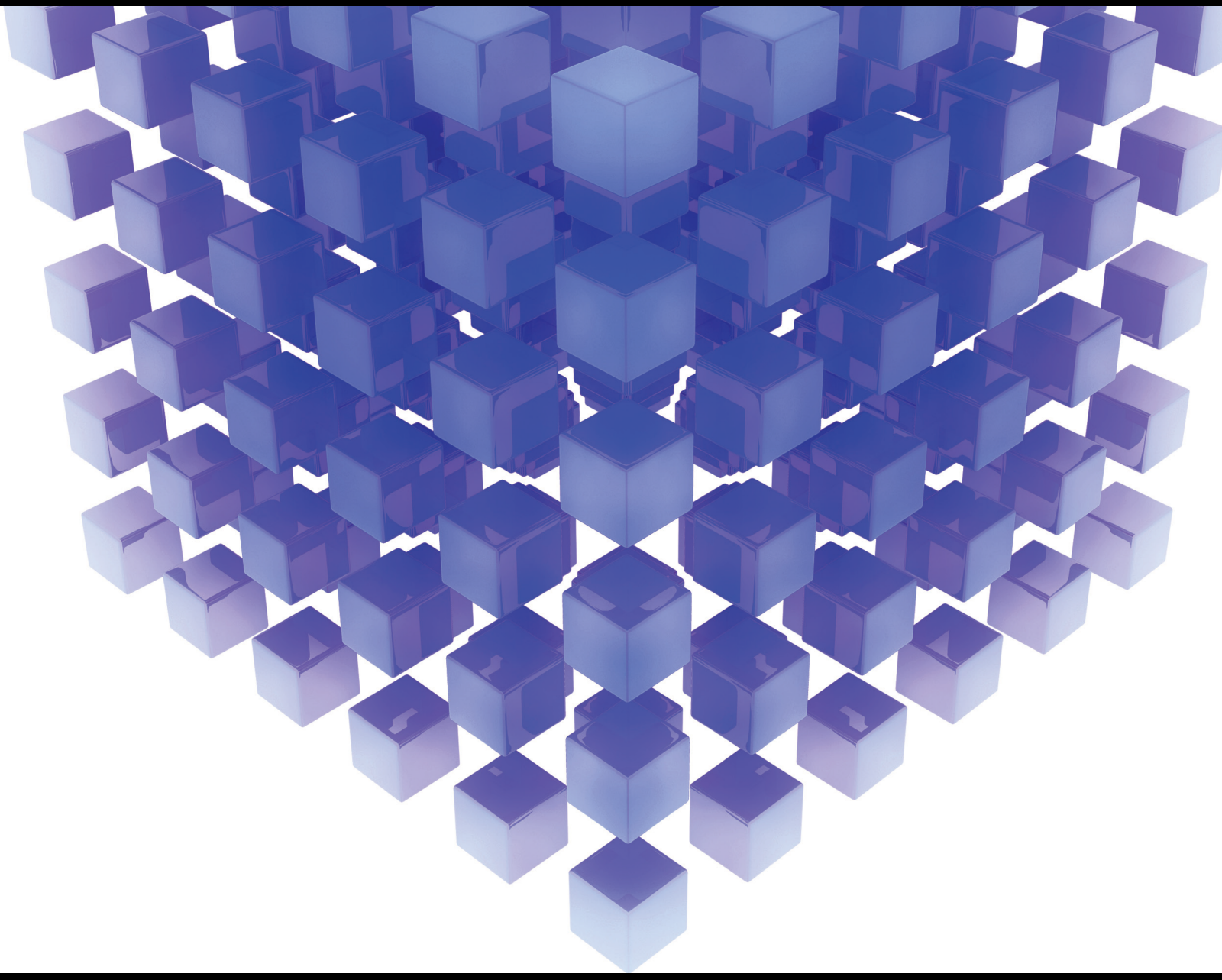
Mathematical Problems in Engineering

Recent Advances in Optimisation Theory, Methods, and Applications in Science and Engineering

Lead Guest Editor: Guoqiang Wang
Guest Editors: Jiyuan Tao, Goran Lesaja, and Mohamed El Ghami



Recent Advances in Optimisation Theory, Methods, and Applications in Science and Engineering

Mathematical Problems in Engineering

Recent Advances in Optimisation Theory, Methods, and Applications in Science and Engineering

Lead Guest Editor: Guoqiang Wang
Guest Editors: Jiyuan Tao, Goran Lesaja, and Mohamed El Ghami

Contents

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9234084, 10 pages
https://doi.org/10.1155/2021/9234084

Research Article

Application of High-Dimensional Outlier Mining Based on the Maximum Frequent Pattern Factor in Intrusion Detection

Limin Shen,[1] **Zhongkui Sun**,[1,2] **Lei Chen**,[3] **and Jiayin Feng**[1]

[1]*School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China*
[2]*Qinggong College, North China University of Science and Technology, Tangshan 063000, China*
[3]*Graduate School, North China University of Science and Technology, Tangshan 063000, China*

Correspondence should be addressed to Zhongkui Sun; sunzhk7965@stumail.ysu.edu.cn

Received 12 August 2020; Revised 1 May 2021; Accepted 14 June 2021; Published 22 June 2021

Academic Editor: Guoqiang Wang

As the Internet applications are growing rapidly, the intrusion detection system is widely used to detect network intrusion effectively. Aiming at the high-dimensional characteristics of data in the intrusion detection system, but the traditional frequent-pattern-based outlier mining algorithm has the problems of difficulty in obtaining complete frequent patterns and high time complexity, the outlier set is further analysed to get the attack pattern of intrusion detection. The NSL-KDD dataset and UNSW-NB15 dataset are used for evaluating the proposed approach by conducting some experiments. The experiment results show that the method has good performance in detection rate, false alarm rate, and recall rate and effectively reduces the time complexity.

1. Introduction

1.1. Intrusion Detection System. With the rapid development of modern information technology, network security has become the focus of attention. How to effectively detect the types of intrusion attacks, as well as the security of the early warning and protection system, has become one of the research directions of network security. Intrusion detection systems (IDSs) are most widely used in the world for identifying and detecting the intruders in computer networks, Internet, and cloud networks. The intrusion detection system analyses the network data collected by the computer system and the key points in the network, so as to find out the behaviour of violating the security policy and the traces of attacks and monitor and detect the network intruders. The IDS can be used to detect different types of attacks on the network, but the traditional firewall cannot perform these attacks well.

Generally, the intrusion detection system can be roughly divided into two categories according to its detection methods, namely, an anomaly detection system and detection system. Anomaly detection is also known as behaviour-based system detection, which detects the abnormal behaviour of the system to discover intrusion behaviour. Misuse detection is a knowledge-based detection or feature-based detection technology, whose premise is that intrusion behaviour and normal network access have different data characteristics. The intrusion detection system is divided into two stages, namely, the preprocessing stage and intrusion detection stage. By developing the intrusion detection system, the intrusion behaviour can be identified effectively.

1.2. Outlier Detection. Outlier mining is an important research direction in the field of data mining. Outlier data do not conform to the general rules of data and are not consistent with other parts of the data. It is those small-scale objects that are far away from other objects in the dataset. Although outlier data are "abnormal data" which are inconsistent with normal data, outlier detection can provide important information in some applications.

There are many reasons for outliers. Generally speaking, they can be divided into two situations: first, they are indeed caused by human or detection equipment errors; second, they are caused by the nature of things themselves, and they are the data reflection of the real nature of things. The outlier

analysed in this paper belongs to the second case. The outlier data generated by human operation are significantly different from the normal network behaviour, in order to find the real potential valuable knowledge through outlier mining.

In the real network activities, most of the network behaviours are normal, the intrusion behaviour can be regarded as the abnormal phenomenon of the amount of data far less than the normal behaviour, and the data corresponding to the normal behaviour and the intrusion behaviour have different data characteristics. Based on the characteristics of intrusion behaviour data, intrusion behaviour can be regarded as "outlier" data [1].

1.3. Association Rule Mining. Association rule mining, as an important part of data mining, has been a hot research topic. Association rules are a collection of items in the database that exceed the specified minimum support and minimum confidence. Association rules are usually expressed as $X \Rightarrow Y$, support $= s$, and confidence $= c$, in which X is the precondition of the rule, Y is the conclusion of the rule, the support s represents the frequency of the rule, and the confidence c represents the strength of the rule.

The goal of association rule mining is to find out all the strong association rules. The mining process is divided into two steps:

Step 1: all rules that are not less than the minimum support threshold s are found, i.e., all frequent patterns

Step 2: by setting the confidence threshold c, the conversion rule is used to filter out the set of items less than the minimum confidence c, and the corresponding association rules are obtained

In this paper, it is only needed to get the maximum frequent patterns based on frequent pattern, so it is only needed to complete Step 1 to get the frequent pattern.

1.4. Maximum Frequent Pattern. If the maximum frequent pattern needs to be explained, the concept of supersets must be introduced first, which is defined as follows: if every element in set S_2 is in set S_1 and set S_1 may contain elements that are not in S_2, then set S_1 is a superset of set S_2. If set S_1 is a superset of set S_2, then set S_2 is a true subset of set S_1, and vice versa.

With the superset, the maximum frequent pattern is defined as follows: if all supersets of frequent pattern X are nonfrequent patterns, then X is called as a maximum frequent pattern.

With the increasing number and dimension of collected data in the intrusion detection system, researchers have proposed a variety of typical high-dimensional outlier mining algorithms for the complexity, sparsity, and diversity of high-dimensional data. Among them, outlier mining based on frequent pattern is widely used in intrusion detection because of its easy-to-understand nature and low time complexity. On the basis of frequent-pattern-based outlier mining algorithm, using the concept of maximum frequent pattern in association rules, an improved high-dimensional outlier mining algorithm based on the maximum frequent pattern is proposed in this paper. The algorithm transforms frequent pattern mining into maximum frequent pattern mining. On the premise of good detection performance, the time complexity is reduced.

2. Literature Survey

In the real network, the data are high dimensional in the intrusion detection system. Some researchers proposed the means to reduce the dimension of high-dimensional data with the way of feature extraction or feature selection and then analysed the processed data with the traditional data mining methods.

Ganapathy [2] proposed an intelligent algorithm for feature selection and classification to design an effective intrusion detection system, which can be used to provide security to networks effectively.

Tian et al. [3] proposed a hierarchical outlier detection model based on PCA, an anomaly data model based on PCA was established based on normal data to filter data firstly, and then, the abnormal data types were analysed to detect both anomaly and misuse attack.

Zyad et al. [4] proposed a way to use the trimmed average vector to estimate the average vector on the basis of PCA, so as to make the trimmed PCA have better robustness.

To solve the problem of high-dimensional data in IDS, Riyaz and Ganapathy [5] proposed a new fuzzy rule and information gain ratio-based feature selection algorithm (FRFSA), and the existing classifiers called SVM and LSSVM were used for effective classification. The experimental result shows that the proposed work exceeds the performance measure when compared to the existing algorithms on classification for feature selection.

Nancy et al. [6] proposed a dynamic recursive feature selection algorithm for feature selection and then used an intelligent fuzzy temporal decision tree algorithm to effectively detect intruders, which can effectively reduce the false positive rate, energy consumption, and delay of the system.

The method of dimension reduction can eliminate some features and reduce the time complexity, but each feature represents a different outlier value. If the features are selected incorrectly, it will get the wrong outlier value, which will produce an approximate result that is not suitable for future calculation [7]. The complexity, sparsity, and diversity of high-dimensional data restrict the traditional mining algorithm. When dealing with high-dimensional data, data mining algorithms suitable for low-dimensional data usually encounter the problems of algorithm efficiency reduction and the traditional definition based on distance and density is invalid, which reduces the accuracy of intrusion detection [8].

Researchers have proposed intrusion detection methods for high-dimensional data. Zhang et al. [9] proposed SPOT technology for anomaly detection in a high-dimensional data network data stream, which has good detection effect.

Prajapati and Bhartiya [10] proposed a nearest neighbour search algorithm based on the advantages of K-mean algorithm and fuzzy C-mean (FCM) algorithm to solve the problem of uneven data and rigid clustering in high-dimensional data, which can realize nearest neighbour search in a shorter time.

In general, the "attack" data in intrusion behaviour are regarded as abnormal data, and outlier mining is to mine those abnormal data which deviate from normal behaviour in large-scale data, so outlier mining is very important for analysing intrusion behaviour. For high-dimensional outlier mining, researchers have proposed several typical mining algorithms: outlier mining algorithm based on spatial projection [11, 12], outlier mining algorithm based on a hypergraph model [13, 14], and outlier mining algorithm based on frequent patterns. The outlier mining algorithm based on frequent patterns is simple, easy to understand, and has lower time complexity than the previous two algorithms, so researchers have conducted extensive research.

In the early stage, He et al. [15] proposed an outlier mining algorithm based on frequent patterns (FindFPOF) and proposed a measurement factor of frequent pattern outlier factor (FPOF). It is believed that the less frequent the patterns contained in a data record, the more likely they would be an outlier, so outliers could be found by calculating the frequent pattern factor of each data.

Zhou [16] proposed a new metric called weighted frequent pattern outlier factor for categorical data streams based on FindFPOF and proposed a fast outlier detection method for high-dimensional categorical data streams based on frequent pattern (FODFP-Stream), which has good applicability and validity.

Wang and Tang [17] proposed an algorithm based on frequent patterns-NFPOF, which further accurately locates abnormal properties of each outlier data through the related attributes of frequent patterns.

Yuan et al. [18] proposed a weighted frequent-pattern-based outlier (WFP-Outlier) to solve the problem whose weights seriously affect outlier detection results, which can find implicit outliers from weighted data streams.

To solve the problem of being incapable of detecting new type of attacks, Jaisankar [19] proposed a new intelligent-agent-based IDS using Fuzzy rough-set-based outlier detection and Fuzzy rough-set-based SVM. The system adopted Fuzzy rough-based SVM in our system to classify and detect anomalies efficiently. The experimental result shows that the proposed intelligent-agent-based model improves the overall accuracy and reduces the false alarm rate.

In order to solve the problem of high false positives, Ganapathy [20] proposed a new intrusion detection model using a new Weighted-Distance-Based Outlier Detection (WDBOD) algorithm and an Enhanced Multiclass Support Vector Machine algorithm, which has low false alarm rate and high accuracy.

Combined with attribute selection, outlier detection, and the enhanced multiclass support vector machine classification method, Ganapathy et al. [21] proposed a new intelligent-agent-based intrusion detection model for mobile ad hoc networks. Using the proposed Intelligent Agent Weighted Distance Outlier Detection algorithm and Intelligent-Agent-based Enhanced Multiclass Support Vector Machine algorithm, the proposed model can detect anomalies with low false alarm rate and high accuracy.

To sum up, high-dimensional outlier mining based on frequent patterns plays a very important role in intrusion detection, but there are two problems in the algorithms based on frequent patterns. First, it needs to mine the complete frequent patterns in the dataset, but it is very difficult to find the complete set of frequent patterns in high-dimensional data. Second, the time complexity of mining algorithm for frequent patterns is exponentially related to the dimension of data, the higher the dimension, the greater the time complexity. High-dimensional outlier mining algorithm based on frequent patterns has the problems of difficulty in obtaining complete frequent patterns and high time complexity. So, a high-dimensional outlier mining algorithm based on the maximum frequent pattern factor is proposed in this paper using the concept of maximum frequent pattern factor in association rules. Also, the algorithm is applied in intrusion detection, which reduces the time complexity on the premise of ensuring good detection performance.

3. Proposed Work

3.1. Relevant Theories. We let $D = \{t_l, t_2, \ldots, t_n\}$ be a dataset containing n network behaviour records t, and t_k is called a transaction. Also, $I = \{i_l, i_2, \ldots, i_p\}$ is the collection of all attributes in the network behaviour record, and i_m is called an item.

Definition 1. Itemset: any subset X of I is called the itemset of D. We let t_k be a transaction of D, and X is a itemset of D; if $X \subseteq t_k$, then the itemset D is contained in the transaction t_k.

Definition 2. Support: the support number of itemset X is represented as the number of transactions that contain itemset X in dataset D and is recorded as X. The support of itemset X is recorded as

$$\text{support}(X) = \frac{X}{D} \times 100\%, \tag{1}$$

where D is the total number of transactions in dataset D.

Definition 3. Frequent pattern: if the support (X) is not less than the minimum support (MinSP) which is specified by the user, then X is a frequent pattern; otherwise, it is an infrequent pattern.

Theorem 1. *X, Y are set as itemsets in dataset D; then,*

(1) If $X \subseteq Y$, then support $(X) \geq$ support(Y)

(2) If $X \subseteq Y$ and X is not a frequent pattern, then Y is not a frequent pattern

(3) If $X \subseteq Y$ and Y is a frequent pattern, then X is a frequent pattern

Y is set as a maximum frequent pattern because $X \subseteq Y$, and Y must be a frequent pattern; it can be seen from Theorem 1 that X must be a frequent pattern, that is to say, all frequent patterns have been implied in the maximum frequent patterns. Therefore, the problem that the complete set of frequent patterns must be found in the outlier mining algorithm based on frequent patterns can be transformed into finding the maximum frequent patterns. It not only solves the difficulty of finding the complete frequent pattern sets but also greatly reduces the number of frequent patterns n, thus reducing the time complexity of the algorithm.

3.2. Data Discretization. The data types of attributes in a dataset can be divided into textual data and numerical data, and numerical data also can be divided into discrete data and continuous data. The data type in outlier mining based on maximum frequent patterns must be discrete data, so it is necessary that continuous attributes are converted to reliable accurate data suitable for data mining by data discretization.

The discretization of numerical attribute is to divide the continuous data into a number of finite discretization intervals. The usual discretization methods include the equal-width method, the equal-frequency method, and the method based on clustering. Clustering is an unsupervised algorithm; according to the distribution characteristics of data to determine how to divide the interval of attribute values, as far as possible to reduce manual intervention, it has been widely used in practice. After clustering, the objects in the same clustering pattern have a high similarity and are quite different from the objects that do not belong to the same clustering pattern, and data in a same clustering pattern are often treated as a whole in many practical applications. In order to minimize the intervention of human factors, the method based on clustering is adopted for data discretization in this paper.

The discretization method based on clustering has two steps:

(1) Continuous attributes are clustered by the clustering algorithm

(2) Patterns obtained by clustering are processed, and continuous attribute values in the same clustering pattern are uniformly marked as one value

Among them, clustering is the key step in discretization. K-means is a classical clustering algorithm based on partition, which has good effect and is widely used in practice. However, K-means algorithm is very sensitive to the number of clustering K and the selection of initial clustering centre.

For the sensitive problem of K value, the elbow method can be used to determine the optimal K value because K value is not fixed and unique in the process of discretization. The core idea of the elbow method is when K is less than the optimal number of clustering, an increase in K value will greatly increase the degree of aggregation of each clustering, so the decrease range of SSE will be very large. When K reaches the true number of clustering, the return of aggregation degree obtained by an increase in K will decrease rapidly, so the decrease degree of SSE will decrease sharply, and if K value is increased continuously, the change of SSE will tend to be gentle, that is to say, the relationship graph between SSE and K is the shape of an elbow, and the corresponding K value of this elbow is the optimal number of clusters.

The square sum of error (SSE) of the core index of the elbow method is defined as

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{p \in C_i} \left| p - m_i \right|^2, \tag{2}$$

where C_i: the i_{th} clustering, p: sample points in C_i, m_i: the centroid of C_i (mean value of all samples in C_i), and SSE: clustering error of all samples, representing the quality of the clustering effect.

For the sensitive problem of the selection of an initial cluster centre, the maximum distance method is used to select K samples as the initial centre points based on the fact that the farthest sample points are most unlikely to be divided into the same cluster.

3.3. The Proposed Algorithm. The concept of maximum frequent pattern factor (MFPOF) is proposed based on the frequent pattern factor (FPOF) in FindFPOF algorithm.

Definition 6. Maximum frequent pattern factor (MFPOF): MFPS (D, MinSP) is the maximum frequent pattern sets in dataset D that meets a given minimum support threshold. The MFPOF of each network behaviour record t is defined as

$$\mathrm{MFPOF}(t) = \frac{\sum_{X \sqsubseteq t,\, X \in \mathrm{MFPS}(D, \mathrm{MinSP})} \mathrm{support}(X)}{\| \mathrm{MFPs}(D, \mathrm{MinSP}) \|}, \tag{3}$$

where $\| \mathrm{MFPs}(D, \mathrm{MinSP}) \|$ is the number of the maximum frequent patterns in frequent patterns and the support(X) is the support of a maximum frequent pattern X.

The description of the high-dimensional outlier mining algorithm based on maximum frequent patterns (MFPOF-OM) is shown as Algorithm 1.

3.4. Automatically Constructing Intrusion Detection Patterns Based on Association. Association analysis can automatically discover the data characteristics of network behaviour. The maximum frequent patterns generated by association analysis can reflect the maximum common characteristics of network behaviour data, which are expressed by the attribute values of network behaviour data. So, these attribute values can be used to build intrusion detection patterns with strong classification ability [22].

Taking the outlier dataset obtained by MFPOF-OM algorithm as input and setting a minimum support threshold, the maximum frequent patterns of the outlier dataset can be obtained referring to Step 1–3 of Algorithm 1, which are the intrusion detection patterns of network attack.

3.5. System Architecture. According to the abovementioned analysis, the architecture of the system proposed in this work consists of six major modules such as data preprocessing, an

```
Input: D//network behaviour dataset
   MinSP//minimum support threshold
   k//number of outliers threshold
Output: k network behaviour outlier data records
Begin
   // Stpe 1–3: mining the maximum frequent item sets based on PF-Tree Algorithm
   Step 1: To D, the HeaderTable (D) is generated to satisfy the MinSP;//Calculating the header table of PF-tree
   Step 2: To D, the frequent item set tree is generated to satisfy the given MinSP by using the PF-Tree Algorithm, and denoted as: T;//
Obtains frequent item set tree according to the PF-Tree algorithm
   Step 3: Obtains maximum frequents item sets based on an improved PF-Tree, and obtains MFPs (D, MinSP) and support (X)//
Obtains maximum frequents item sets
   //Stpe 4–7: Mine k outliers data with minimum MFPOF value based on the obtained MFPs
   Step 4: foreach t in D
   According to formula (3), calculates the maximum frequent patterns factor of each record t: MFPOF(t);
   end foreach//Calculating maximum frequent factor of each transaction t
   Step 5: Obtains a MFPOF value of each network behaviour records t;
   Step 6: For all t, they are sorted in ascending order according to MFPOF (t);
   Step 7: Return the first k network behaviour record with the minimum MFPOF value, and they are k outlier data in the network
behaviour data.
End
```

Algorithm 1: MFPOF-OM algorithm.

outlier mining module, constructing intrusion detection patterns, attack patterns base, pattern match, and an alarm system, as shown in Figure 1.

The data preprocessing module is for performing preprocessing activities, but its main function is to discretize the data and make it suitable for the proposed algorithm. The outlier mining module is used to obtain the outlier data by the proposed algorithm. On the basis of acquiring outlier data, an intrusion detection pattern module is used to obtain intrusion detection patterns, so as to construct the attack pattern library module. The pattern match module is used to match the testing data with the attack rule base. If the match is successful, it indicates that there is an intrusion attack and transfers to the alarm module to trigger the alarm.

4. Results and Discussion

4.1. Dataset and Experimental Environment. The specifications of the hosts adopted in the experiments are Core Intel Core i5-6300HQ, 2.3 GHz CPU, 16 GB RAM, and Windows 7. The proposed method is verified in MATLAB 2012. The NSL-KDD dataset [23] and UNSW-NB 15 dataset [24] are used as the experimental datasets to verify the proposed method in this paper.

First, the experimental results of the proposed algorithm are analysed in the NSL-KDD dataset, and then, the proposed algorithm is compared with other researchers' algorithms to verify the effectiveness it; lastly, the experimental results in the NSL-KDD dataset and UNSW-NB 15 dataset are compared to verify the applicability of the proposed algorithm.

The NSL-KDD dataset is an effective benchmark dataset to help researchers compare different intrusion detection methods. There are 125,973 connection records in the NSL-KDD dataset. Each connection record is described by 41 attributes about the network packet, network traffic, host traffic, and content information. The 22 categories of attacks are from the following four classes: DoS, R2L, U2R, and Probing. Also, the 20th attribute (num_outbound_files) can be deleted because its attribute value is all 0, so its information entropy is 0 according to information theory.

The raw network packets of the UNSW-NB15 dataset are created for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours. It is suitable for researchers to study the intrusion detection system. There are 175,341 records in the training set and 82,332 records in the testing set. This dataset has totally 49 features with the class label and 9 families of attacks, namely, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms.

The NSL-KDD dataset is a factual benchmark in the field of network intrusion detection, which lays a foundation for the research of network intrusion detection based on computational intelligence. First, the NSL-KDD dataset eliminates duplicate records and classifiers that prefer more duplicate records. Second, it eliminates the imbalance between the number of records and reduces the false positive rate. Therefore, although the NSL-KDD dataset is older, it is widely used to evaluate the performance of the IDS. The UNSW_NB15 dataset is a comprehensive network attack traffic dataset, which combines the real normal network traffic attack activities and modern network traffic comprehensive attack activities and can better reflect the real environment of the network, so it is widely used in abnormal intrusion detection [25, 26].

The proposed algorithm needs to mine the maximum frequent pattern, which requires that the data type must be discrete. Taking the NSL-KDD dataset as an example, the dataset values' processing is introduced, which is suitable for the proposed algorithm. According to the analysis of the

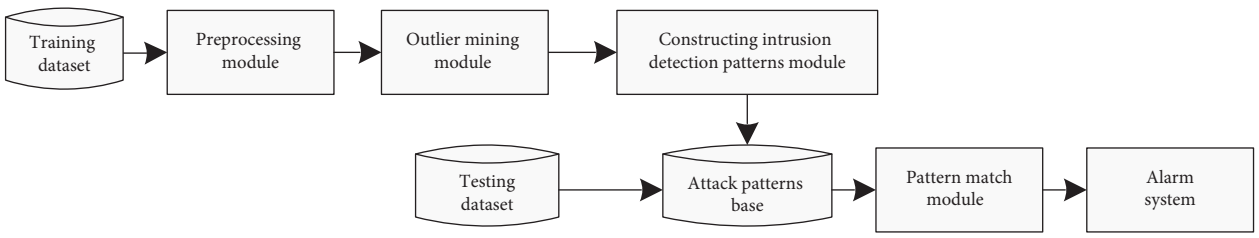


Figure 1: System architecture.

NSL-KDD dataset, the attribute data type of the dataset can be divided into the text type and numerical type, and the numerical type can be divided into the discrete type and continuous type. The types of data are shown in Table 1 for the text-type and numerical discrete-type data which have met the data requirements. However, the continuous numerical data represented by columns 1, 5, and 6 are discretized using the discretization algorithm given in Section 3.2 and transformed into reliable and accurate data suitable for data mining.

4.2. Experiments in the NSL-KDD Dataset. Experiment A: the experimental results of the proposed algorithm in the NSL-KDD dataset are analysed in the experiment. The accuracy, false positive rate, and complexity analysis are used as the performance evaluation criteria to determine the results. Four groups of sample data were extracted from the dataset: Normal + DoS, Normal + Probing, Normal + R2L, and Normal + U2R.

4.2.1. Experiment Results of Four Network Attack Patterns. By comparing the detection rate and false positive rate under different *MinSP* thresholds of four groups of sample data, Normal + DoS, Normal + Probing, Normal + R2L, and Normal + U2R, the detection effect of the proposed algorithm is illustrated, and then, the feasibility of the proposed algorithm is verified. The experimental results of DoS, Probing, R2L, and U2R intrusion detection patterns obtained from the analysis of four groups of sample data are shown in Figure 2.

Probing attack detection patterns are taken as an example for data analysis. The Normal + Probing sample set contains 62000 pieces of data, the threshold value of MinSP is different, and the detection patterns are also different in the experiment. The experimental results are shown in Figure 2(b), which shows the detection patterns acquired under the MinSP thresholds of 58500, 59000, and 60000 and uses the acquired Probing detection patterns to detect five data types (DoS, Probing, R2L, U2R attack data, and Normal data), respectively. It is found that when the threshold value is 59000, the accuracy of Probing detection patterns to Probing data is 88%, and the false alarm rate is 2% to Normal data, 4% to DoS, 1% to R2L, and 10% to U2R data. When the threshold values are 58000 and 60000, the results are as shown in Figure 2(b) and will not be described one by one.

By comparing the four intrusion detection attack modes in Figure 2, it is found that the accuracy will be better when the minimum support threshold is larger, and the detection error of other data is basically the same, although the size varies. It is determined by the characteristics of outlier mining. The larger the threshold is, the fewer the number of outliers is, which can better reflect the characteristics of attack-type data. Of course, the threshold should not be too large, and the accuracy will be reduced if the threshold is too large. Through the comprehensive analysis of detection rate and false detection rate under multiple thresholds, the intrusion detection mode with the best comprehensive detection result is selected as the acquired intrusion detection mode, and the threshold value at this time is taken as the acquired intrusion detection pattern threshold: the threshold of DoS attack is 59100, the threshold of Probing attack is 59000, the threshold of R2L attack is 59600, and the threshold of U2R attack is 59500. The evaluation parameters are shown in Table 2.

Table 1: NSL-KDD dataset attribute data types.

Attribute types	Column
Text type	2, 3, and 4
Numerical discrete type	7, 12, 14, 15, 21, and 22
Continuous numerical data	1, 5, 6, and other columns

Comparing the four subgraphs in Figure 2, it is found that U2R-type data have the highest detection errors in DoS, Probing, and R2L attack intrusion detection patterns, which are 4%, 10%, and 33%, respectively, and compared with the other three attack intrusion detection patterns, the accuracy of U2R attack intrusion detection mode is relatively low, only 87%, which is determined by the number of U2R, only 52 pieces of U2R data in the NSL-KDD dataset, so data mining cannot fully discover its data characteristics, resulting in incomplete detection performance.

Comparing Figure 2(c) with Figure 2(d), it is found that there are higher errors in the detection of U2R data by using R2L attack intrusion detection patterns and R2L data by using U2R attack intrusion detection patterns, which shows that R2L-type data and U2R-type data have higher data similarity compared with other three types of data, which is consistent with the characteristics of two kinds of network attacks in reality.

4.2.2. Complexity Analysis. In this section, the complexity of 4 groups of sample data, Normal + DoS, Normal + Probing, Normal + R2L, and Normal + U2R, will be analysed. The FindFPOF algorithm based on frequent patterns and other outlier mining algorithms based on weighted frequent patterns need to mine frequent patterns first, and the time complexity is similar. Here, FindFPOF algorithm is taken as an example to illustrate.

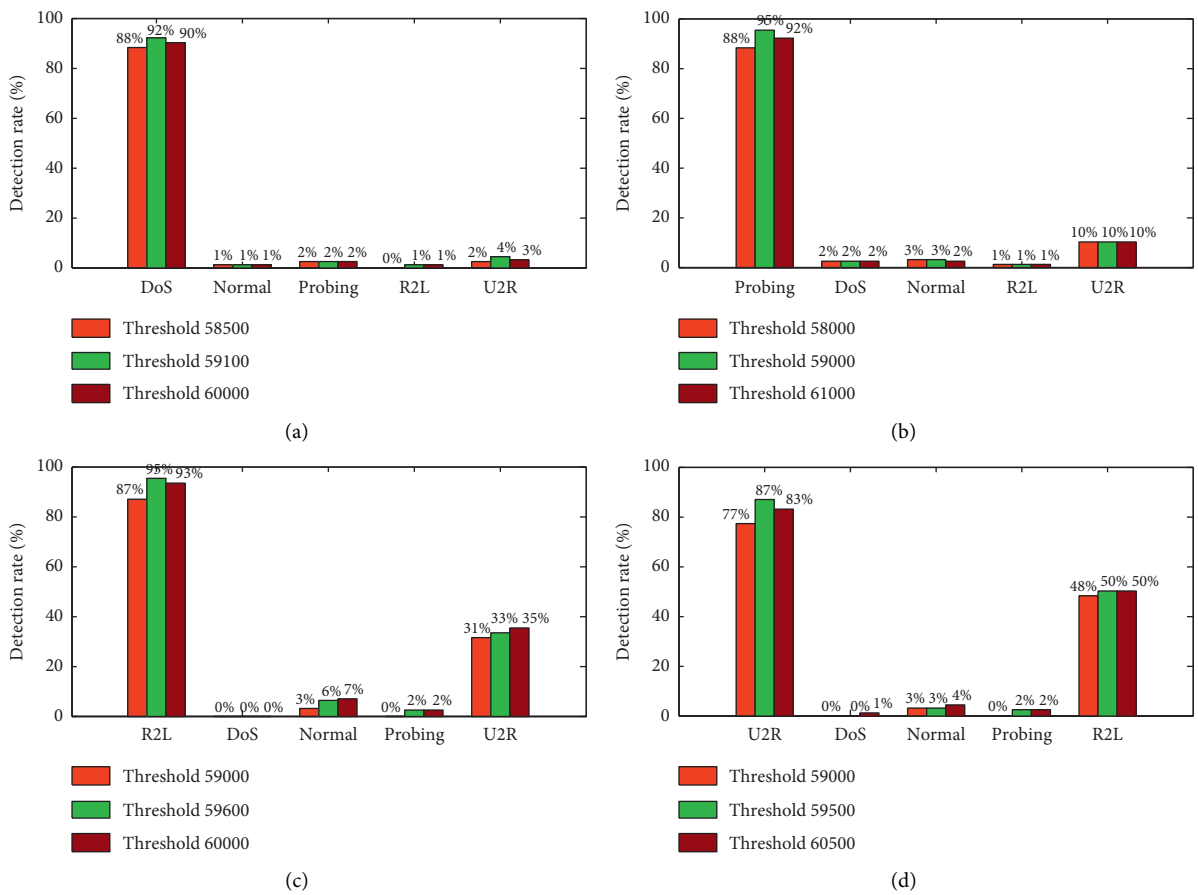


Figure 2: Test results of four network attacks. (a) Test results of DoS misuse detection patterns. (b) Test results of Probing misuse detection patterns. (c) Test results of R2L misuse detection patterns. (d) Test results of U2R attack misuse detection patterns.

The total time complexity of FindFPOF algorithm is $O(m^2 + m*n + m*logm)$, where m is the amount of data and n is the amount of frequent patterns.

The MFPF-OM algorithm has three steps: (1) mining maximum frequent patterns from the dataset, the time complexity is $O(m^2)$; (2) calculating the MFPOF(t) of each network behaviour record, the time complexity is $O(m*l)$; and (3) discovering K network behaviour outliers, the time complexity is $O(m*logm)$. Therefore, the time complexity from the abovementioned three steps is proved as follows: $T(\text{MFPOF-OM}) = O(m^2 + m*l + m*logm)$, where m is the number of data and l is the number of maximum frequent patterns.

The number of frequent patterns (n) in FindFPOF algorithm and the number of maximum frequent patterns (l) in MFPF-OM algorithm for 4 groups of sample are shown in Table 3.

For massive data, the value of m is large enough, and in theory, the time complexity of the two algorithms can be simplified to $O(m^2)$. But in practice, when the value of m is not large enough, the proposed algorithm only needs to mine the maximum frequent patterns in Step 3, and $l \ll n$, as shown in Table 3, so MFPOF-OM algorithm has a better time complexity than the FindFPOF algorithm when calculating MFPOF (t) in Step 4 of the algorithm.

4.3. Comparative Experiments between the Proposed Algorithm and Other Algorithms. Experiment B: in order to verify the accuracy of the proposed method, it is compared with the SVM method, Intelligent DT method [6], LSSVM + FRFSA method [5], and Outlier Detection + EMSVW method [20]. The accuracy is used as the performance evaluation criteria to determine the results. The evaluation parameters are shown in Table 4.

The results are shown in Figure 3, in which M1 represents the SVM method, M2 represents the Intelligent DT method, m3 represents the LSSVM + FRFSA method, M4 represents the Outlier Detection + EMSVW method, and M5 represents the proposed method in this paper. The results show that the MFPOF-OM method is very close to the other methods in accuracy of Probing and DoS, but slightly inferior. However, it has a great advantage in the accuracy of R2L and U2R, which shows that the improved dimensional outlier mining method has good characteristics in dealing with outlier data because of the small amount of R2L and U2R attack data in the NSL-KDD dataset. The accuracy data of R2L and U2R are empty in Figure 3 because there are no relevant data in [20]. The overall performance analysis shows that the performance of the proposed method is reliable, can effectively detect the intrusion behaviour in network data, and can meet the actual operation requirements.

Table 2: The result of two mining algorithms.

Sample set (sample size)	Threshold value	Accuracy (%)	False positive rate (%)				
			Normal	DoS	Probing	R2L	U2R
Normal + DoS (63000)	58500	88	1	Null	2	0	2
	59100	92	1	Null	2	1	4
	60000	90	1	Null	2	1	3
Normal + Probing (62000)	58000	88	2	3	Null	1	10
	59000	95	2	4	Null	1	10
	61000	92	2	2	Null	1	10
Normal + R2L (60900)	59000	87	3	0	2	Null	31
	59600	95	6	0	2	Null	33
	60000	93	7	0	2	Null	35
Normal + U2R (60052)	59000	77	3	0	0	48	Null
	59500	87	3	0	2	50	Null
	60500	83	4	1	2	50	Null

Table 3: The result of two mining algorithms.

Sample dataset	Number of samples (m)	Number of FP (n)	Number of MFP(l)
Normal + DoS	63000	23	4
Normal + Probing	62000	19	1
Normal + R2L	60900	21	3
Normal + U2R	60052	23	2

Table 4: Comparison of detection rates of different algorithms.

	SVM	Intelligent DT	LSSVM + FRFSA	Detection + EMSVW	Proposed method
Probing	95.42	99.59	92	99.1	95
DoS	94.29	99.2	95	99.2	92
R2L	45.34	50.88	38	Null	95
U2R	31.34	35.88	38	Null	87

4.4. Comparative Experiments between the NSL-KDD Dataset and UNSW-NB15 Dataset. Experiment C: in this experiment, the proposed method is tested and compared in the NSL-KDD dataset and UNSW-NB15 dataset, and the performance of the proposed algorithm is estimated by using the performance metrics, namely, precision, recall, and F1-measure and ROC. The two datasets have different attack patterns and data characteristics, so it is impossible to compare each pattern separately, and only the overall performance index is analysed in two datasets in this paper. The overall performances of precision, recall, and F1-measure in the two databases are shown in Table 5. Figure 4 shows the comparison results of precision, recall, and F1-measure in two different databases.

Figure 5 shows the ROC curves in two different databases. It is found that although the detection results of the UNSW-NB15 dataset are better than those of the NSL-KDD dataset in some values, the detection results of the NSL-KDD dataset are generally better than those of the UNSW-NB15 dataset from the whole ROC curve.

By comprehensively comparing the performance indexes in Figures 4 and 5, it is found that the proposed method's technique achieves better performances for the NSL-KDD dataset. The reason is that some malicious records in the UNSW-NB15 one are not high because of the lower variances between them and normal records, and the data are optimized in the NSL-KDD database, which is more suitable for the detection of malicious records. But on the whole, it shows very good performance in the NSL-KDD dataset and UNSW-NB15 dataset, which proves the effectiveness of the proposed method in high-dimensional anomaly detection.

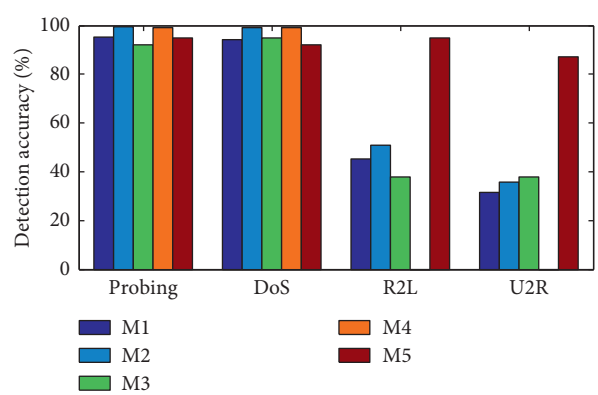


Figure 3: Comparison between other intrusion detection methods and the method proposed.

Table 5: Performance comparison between the two databases.

	Precision (100%)	Recall (100%)	F1-measure (100%)
NSL-KDD	94	91	92
UNSW-NB15	91	89	90

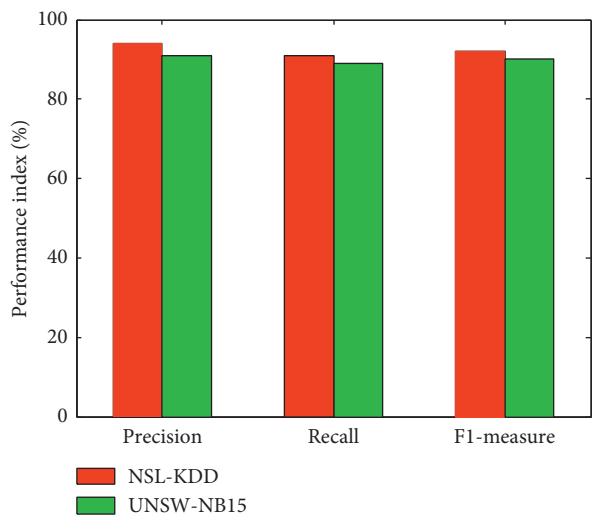


Figure 4: Comparison between the NSL-KDD dataset and UNSW-NB15 dataset.

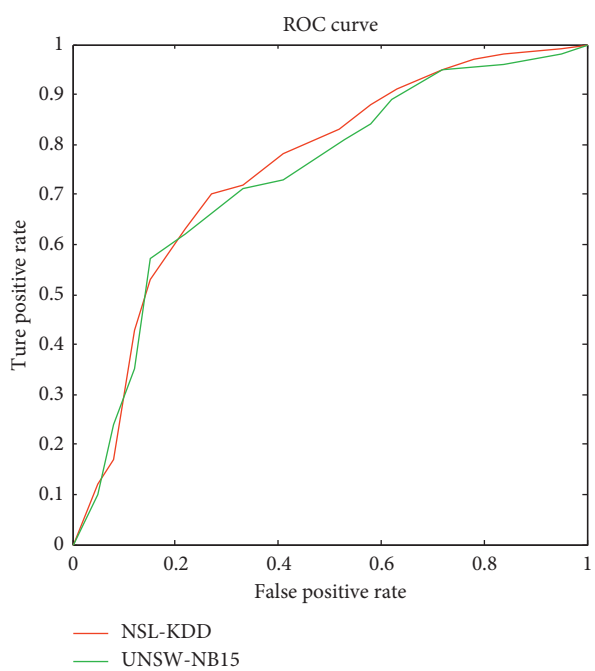


Figure 5: ROC curve of the NSL-KDD dataset and UNSW-NB15 dataset.

5. Conclusions

In this paper, a high-dimensional outlier mining algorithm based on the maximum frequent pattern factor (MFPOF-OM) has been proposed by using the related technology of high-dimensional outlier mining based on frequent patterns. This work has two advantages: first, the MFPOF-OM algorithm only needs to mine the maximum frequent pattern set, which solves the problem of mining completely frequent patterns in frequent pattern outlier algorithm; second, it can greatly reduce the number of maximum frequent patterns, thus reducing the time complexity of the algorithm. Experimental results show that the proposed method is feasible, which can further reduce the time complexity while ensuring the excellent detection performance compared with the contrast algorithms.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61772450) and Hebei Province Natural Science Foundation of China (F2017203307).

References

[1] B. Huang, "Intrusion detection technology based on outlier mining," *Computer Engineering*, vol. 3, pp. 88–90, 2008.

[2] S. Ganapathy, "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, 16 pages, 2013.

[3] B. Tian, K. Merrick, S. Yu, and J. Hu, "A hierarchical pea-based anomaly detection model," in *Proceedings of the 2013 International Conference on Computing, Networking and Communications (ICNC)*, pp. 621–625, IEEE, San Diego, CA, USA, January 2013.

[4] E. Zyad, A. Taha, and B. Mohammed, "Improve R2L attack detection using trimmed PCA," in *Proceedings of the 2019 International Conference on Advanced Communication*

Technologies and Networking (CommNet), pp. 1–5, IEEE, Rabat, Morocco, April 2019.

[5] B. Riyaz and S. Ganapathy, "An intelligent fuzzy rule based feature selection for effective intrusion detection," in *Proceedings of the 2018 International Conference on Recent Trends in Advance Computing (ICRTAC)*, pp. 207–211, IEEE, Chennai, India, September 2018.

[6] P. Nancy, S. Muthurajkumar, S. Ganapathy, S. V. N. Santhosh Kumar, M. Selvi, and K. Arputharaj, "Intrusion detection using dynamic feature selection and fuzzy temporal decision tree classification for wireless sensor networks," *IET Communications*, vol. 14, no. 5, pp. 888–895, 2020.

[7] G. L. Prajapati and R. Bhartiya, "High dimensional nearest neighbor search considering outliers based on fuzzy membership," in *Proceedings of the 2017 Computing Conference*, Bologna, Italy, July 2017.

[8] S. Zhou, *Research on Algorithm of High Dimensional Outlier Detection*, MS thesis, Jiangsu University, Zhenjiang, China, 2007.

[9] J. Zhang, Q. Gao, and H. Wang, "Anomaly detection in high-dimensional network data streams: a case study," in *Proceedings of the 2008 IEEE International Conference on Intelligence and Security Informatics*, pp. 251–253, IEEE, Taipei, Taiwan, June 2008.

[10] G. L. Prajapati and R. Bhartiya, "High dimensional nearest neighbor search considering outliers based on fuzzy membership," in *Proceedings of the 2017 Computing Conference*, pp. 363–371, IEEE, London, UK, July 2017.

[11] P. Guo, J.-y. Dai, and Y.-X. Wang, "Outlier detection in high dimension based on projection," in *Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, pp. 1165–1169, IEEE, Dalian, China, August 2006.

[12] H. Liu, "Efficient outlier detection for high-dimensional data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2451–2461, 2017.

[13] Y. Z. Li, "An improved outlier detection method in high-dimension based on weighted hypergraph," in *Proceedings of the 2009 Second International Symposium on Electronic Commerce and Security*, pp. 159–163, IEEE, Lyon, France, August 2009.

[14] N. Wang, Z. Zhang, X. Zhao, Q. Miao, R. Ji, and Y. Gao, "Exploring high-order correlations for industry anomaly detection," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9682–9691, 2019.

[15] Z. He, X. Xu, Z. Huang, and S. Deng, "FP-outlier: frequent pattern based outlier detection," *Computer Science and Information Systems*, vol. 2, no. 1, pp. 103–118, 2005.

[16] X.-Y. Zhou, "A fast outlier detection algorithm for high dimensional categorical data streams," *Journal of Software*, vol. 18, no. 4, pp. 933–942, 2007.

[17] Q. Wang and R. Tang, "Application of frequent pattern based outlier mining in intrusion detection," *Application Research of Computers*, vol. 30, no. 4, pp. 1208–1211, 2013.

[18] G. Yuan, S. Cai, and S. Hao, "A novel weighted frequent pattern-based outlier detection method applied to data stream," in *Proceedings of the 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 503–510, IEEE, Chengdu, China, April 2019.

[19] N. Jaisankar, "An intelligent agent based intrusion detection system using fuzzy rough set based outlier detection," *Soft Computing Techniques in Vision Science*, Springer, Berlin, Heidelberg, 2012.

[20] S. Ganapathy, "An intelligent intrusion detection system using outlier detection and multiclass SVM," *International Journal on Recent Trends in Engineering & Technology*, vol. 5, no. 1, 1953.

[21] S. Ganapathy, P. Yogesh, and A. Kannan, "Intelligent agent-based intrusion detection system using enhanced multiclass SVM," *Computational Intelligence and Neuroscience*, vol. 2012, Article ID 850259, 2012.

[22] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in *Proceedings of the 1999 IEEE Symposium on Security and Privacy (Cat. No. 99CB36344)*, pp. 120–132, IEEE, Oakland, CA, USA, May 1999.

[23] Canadian Institute for Cybersecurity, "The NSL-KDD dataset," 2020, http://www.unb.ca/cic/datasets/nsl.html.

[24] Unsw.adfa.au, "The UNSW-NB15 dataset," 2020, http://www.cybersecurity.unsw.adfa.edu.au/ADFA%20NB15%20.

[25] N. Moustafa, J. Slay, and G. Creech, "Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks," *IEEE Transactions on Big Data*, vol. 5, no. 4, p. 1, 2017.

[26] N. Moustafa and S. Jill, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proceedings of the 2015 military communications and information systems conference (MilCIS)*, IEEE, Canberra, Australia, November 2015.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6697942, 14 pages
https://doi.org/10.1155/2021/6697942

Research Article

Nonlinear Contour Tracking of a Voice Coil Motors-Driven Dual-Axis Positioning Stage Using Fuzzy Fractional PID Control with Variable Orders

Syuan-Yi Chen [1] **and Meng-Chen Yang** [2]

[1] *Department of Electrical Engineering, National Taiwan Normal University, Taiwan, China*
[2] *Department of Hardware Design, Imagination Broadway Ltd., New Taipei City, China*

Correspondence should be addressed to Syuan-Yi Chen; chensy@ntnu.edu.tw

Received 4 November 2020; Revised 1 December 2020; Accepted 2 March 2021; Published 25 March 2021

Academic Editor: Guoqiang Wang

This study aims to develop a variable-order fuzzy fractional proportional-integral-differential (VOFFPID) control system for controlling the mover position of a newly designed voice coil motors- (VCMs-) driven dual-axis positioning stage. First, the operation principle and dynamics of the stage are analyzed. After that, the design of a fuzzy fractional proportional-integral-differential (FFPID) control system is introduced on the basis of a fractional calculus and fuzzy logic system. With an additional degree of freedom to the control parameters and fuzzy operation, the FFPID control system can upgrade the contour tracking performance of a conventional proportional-integral-differential (PID) control system with respect to the specified dynamics of the stage. Moreover, the VOFFPID control system is designed to further improve the tracking responses of the FFPID control system. In this system, the five control parameters are optimized with the cuckoo search algorithm via an adaptive strategy. Lastly, nominal and payload conditions attributed to two nonlinear contour demands are provided to evaluate the contouring performance of the PID, FFPID, and VOFFPID control systems. The experimental results subjected to different performance measures demonstrate that the proposed VOFFPID controller outperforms PID and FFPID controllers in terms of the designed VCMs-driven dual-axis positioning stage under both conditions.

1. Introduction

Although control engineers prefer a conventional proportional-integral-differential (PID) controller because of its easy implementation, low cost, and uncomplicated structure, they cannot use it to achieve a high-precision control level in a highly nonlinear and disturbed situation. To address this problem, a fractional-order (FO) PID (FPID) control method was developed by adding fractional differential and integral operations. With the consideration of more degrees of freedom for the selection of control parameters, the FPID controller can obtain better control responses and anti-interference characteristic over the integer-order (IO) counterparts because of the additional flexibility to the design of a control system [1]. However, accurately determining numerous control parameters in practical applications is difficult. Therefore, many intelligent strategies were designed for FPID control [2, 3].

The introduction of a fuzzy logic system (FLS) to a PID controller has been widely explored because it provides a flexible and model-free way to determine the PID control parameters through engineering intuitions and experiences [4, 5]. In addition, fuzzy FPID (FFPID) control systems were further developed to enhance the control performances of a FPID controller [1, 6–9]. In the FFPID, the fractional operation of errors introduces an extra degree of flexibility in the input variables of FLS, and it can be tuned similarly to the input-output scaling factors of the FLS to enhance the closed-loop performance. Some experimental results have verified that the FFPID control system outperforms classical PID, fuzzy PID, and FPID control systems because of its FLS and higher degrees of freedom for tuning.

Swarm intelligence algorithms have been widely applied to solve many real-world problems, such as control system design [2], path planning [10], parameter estimation [11], and energy management [12], because these algorithms can obtain a global optimal solution for multidimensional optimization problems by relying on colony behaviors in nature. For example, inspired by the aggressive reproduction behavior of cuckoo bird species, cuckoo search algorithms (CSAs) were developed in [2, 13–16]. In cuckoo reproduction, female cuckoos fly from one nest to another and randomly lay their fertilized eggs inside other host birds' nests instead of building their own nests. Thus, host birds may unknowingly raise these eggs. In general, female cuckoos choose the best nest so that their eggs have the best chance of hatching and creating a new generation. To enhance the hatching chance, some cuckoo birds strategically lay their eggs in a good position or drop the host bird's eggs outside its nest. Some cuckoo species even evolve to produce eggs similar to those of other bird species. However, if an alien egg is found, host birds throw it out or even transfer to a new nest with their own brood elsewhere. In this case, the eggs of cuckoo fail to hatch. In the CSA algorithm, cuckoo birds represent the particles assigned to find the optimal solution, while cuckoo eggs and host birds' eggs represent the new and old solutions for the current iteration process, respectively. If a new solution is better than the old one, the worse one is replaced.

A linear voice coil motor (VCM) is a direct drive and hysteresis-free device, which utilizes a magnetic field generated by a permanent magnet with a coil wire to produce an electric driving force [17, 18]. This device has a compact structure, high acceleration, and no hysteresis features, so it has been extensively used in various small range positioning applications, such as servo valves [17], hard disk drives [19, 20], automatic transmitters [21], autofocus actuators [22], and ultrasound scanners [23], which demand high-precision and high-speed control levels. However, effective controls for this device should be designed because external disturbances and operational changes instantly act on a direct drive system. For instance, an intelligent FO sliding-mode control was proposed to control a linear voice coil actuator for the tracking of a reference trajectory [18]. In this control scheme, a fuzzy neural network was designed to compensate for system uncertainties, thereby reducing the chattering phenomena. Moreover, a coupling controller design was proposed by considering the interaction between a VCM and a piezoactuator of a head positioning control system [20]. In another study [24], a direct amplitude control strategy was developed to improve the amplitude accuracy of a reciprocating rig in a high-frequency band compared with that given by a traditional proportional-integral control strategy.

In the direct drive VCM system, there are no mechanical reduction and transmission components so that the mover is directly coupled to the payload. Compared with the conventional rotary motor using mechanical components to translate the rotary motion into linear motion, direct drive device apparently reduces mechanical loss, system nonlinearities, and backlash [25]. Thus, the control accuracy of the VCM system can be enhanced in practical applications. However, it also loses the advantage of using mechanical components attenuating the effects of system parameter variations and external disturbances. With this structure, the system uncertainties are directly transmitted to the payload and then unavoidably affect the control performance of the payload. On the other hand, any change or disturbance in the payload will be directly reflected back to the VCM. Although many control methods have been proposed to control the single-axis VCM systems [17, 18, 21–24], designing effective and robust control methods to meet high-precision requirements for the multi-axis VCM systems is still required. As a result, this study aims to develop a variable-order FFPID (VOFFPID) control strategy for controlling the mover position of a VCM-based dual-axis positioning stage with a high-precision contouring performance. In the VOFFPID controller, control parameters are self-tuned to deal with system uncertainty so that the trivial trials of control parameters are unnecessary. Furthermore, good stability and robustness during the control process can be ensured. Experiments involving the tracking of two nonlinear contour demands were conducted by using PID, FFPID, and VOFFPID under nominal and payload conditions to demonstrate the different control performance and robustness levels.

From the aforementioned studies, the main academic and industrial contributions of this study are summarized as follows: (i) the new VOFFPID controller that optimizes the conventional FFPID controller online is successfully developed; (ii) the new VCMs-driven dual-axis positioning stage is made with operation and dynamic analyses; (iii) the PID, FFPID, and VOFFPID controllers for the VCMs-driven dual-axis positioning stage control system are successfully implemented; and (iv) the experimental results of the three controllers associated with two nonlinear contour tracking commands under two test conditions are compared. The remaining parts of this study are organized as follows. The operation principle of VCMs-driven dual-axis positioning stage is described in Section 2. The CSA with the adaptive strategy used for optimizing the control parameters of the VOFFPID is presented in Section 3. The designs of contour tracking controllers are introduced in Section 4. The experimental setup and results are discussed in Section 5. The conclusions of the proposed work are provided in Section 6.

2. Operation Principle of the VCMs-Driven Dual-Axis Positioning Stage

A circular moving coil-type single-axis VCM that is composed of a moving coil winding and a stationary permanent magnet within a soft iron shell is utilized in this study as shown in Figure 1. In accordance with the interaction between the permanent magnetic field and a drive current perpendicular to the field, the mover of the VCM moves along the direction of the electric driving force, which can be determined with Fleming's left-hand rule [7]. If the direction of the drive current changes, the moving direction also reverses. Moreover, the generated electric driving force is

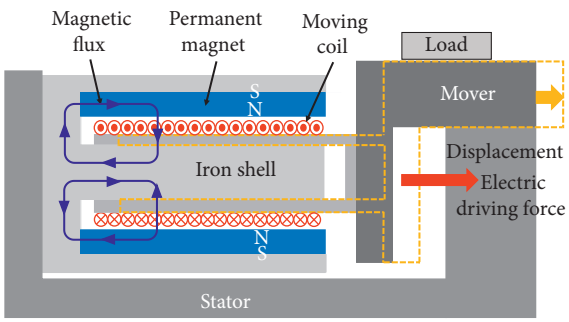


Figure 1: Structure of the VCM.

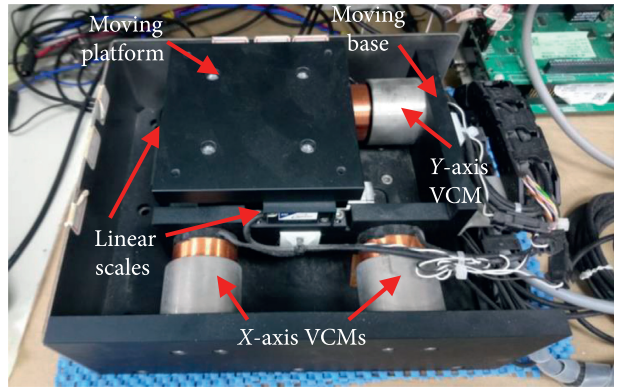


Figure 2: Structure of the VCMs-driven dual-axis positioning stage.

proportional to the product of the permanent magnetic field and the drive current [18].

In this study, a VCMs-driven dual-axis positioning stage is newly designed and implemented as shown in Figure 2. The dimension of the whole stage is 230 mm × 194 mm × 100 mm. It is composed of three VCMs (Akribis, AVM 40-20), namely, a VCM in the Y-axis and two parallel VCMs in the X-axis. A 100 mm^2 moving platform is placed on the mover of the Y-axis VCM, and the stator of the Y-axis VCM is mounted on a moving base. With the design of this stage, two VCMs in the X-axis can generate a stronger electromagnetic force to push the moving platform, moving base, Y-axis VCM, and payload along the X-axis. They can even create rotational motion according to the specified mechanism design and different displacements of X-axis VCMs. Two high-resolution linear scales measure the mover displacements for high-precision and repeatability applications. Specifications of the adopted VCMs are listed in Table 1 [26].

Table 1: Specifications of the adopted VCMs.

Specifications	Value	Unit
Diameter	40	mm
Stroke	20	mm
Force-current coefficient	12.90	N/A
Back electromotive force constant	12.90	V/m/s
Continuous force	9.93	N
Peak force	58.05	N
Continuous current	0.77	A
Peak current	4.5	A
Continuous power	7.17	W
Coil assembly mass	67.0	g
Core assembly mass	226.2	g

The state-space model is given below to describe the dynamics of the VCMs-driven dual-axis positioning stage [17]:

$$\begin{cases} \ddot{x} = -\dfrac{k_{kx}+\Delta k_{kx}}{m_b+m_p+m_l}x - \dfrac{k_{bx}+\Delta k_{bx}}{m_b+m_p+m_l}\dot{x} + \dfrac{c_x k_{ix}+\Delta c_x k_{ix}}{m_b+m_p+m_l}u_x - \dfrac{F_{fx}+F_{dx}}{m_b+m_p+m_l}, \\ \ddot{y} = -\dfrac{k_{ky}+\Delta k_{ky}}{m_p+m_l}y - \dfrac{k_{by}+\Delta k_{by}}{m_p+m_l}\dot{y} + \dfrac{c_y k_{iy}+\Delta c_y k_{iy}}{m_p+m_l}u_y - \dfrac{F_{fy}+F_{dy}}{m_p+m_l}, \end{cases} \tag{1}$$

where x and y are the mover positions in X-axis and Y-axis, respectively; u_x and u_y indicate the control signals of the VCMs; c_x and c_y represent the linear gains of the current amplifiers; k_{ix} and k_{iy} are the force-current coefficients of the VCMs; k_{bx} and k_{by} are the equivalent damping coefficients; k_{kx} and k_{ky} denote the equivalent elastic coefficients; Δk_{kx}, Δk_{ky}, Δk_{bx}, Δk_{by}, $\Delta c_x k_{ix}$, and $\Delta c_y k_{iy}$ represent the unknown parameter variations of k_{kx}, k_{ky}, k_{bx}, k_{by}, $c_x k_{ix}$, and $c_y k_{iy}$, respectively; m_b, m_p, and m_l denote the masses of the moving base, platform, and payload, respectively; F_{fx} and F_{fy} are the friction forces; and F_{dx} and F_{dy} denote the unmodeled system uncertainties, comprising internal cross-coupled interferences and external disturbances. Thus, the dynamic model of the VCMs-driven dual-axis positioning stage can be reexpressed as

$$\begin{cases} \ddot{x} = -\dfrac{k_{kx}}{M_x}x - \dfrac{k_{bx}}{M_x}\dot{x} + \dfrac{c_x k_{ix}}{M_x}u_x - \dfrac{L_x}{M_x}, \ddot{y} = -\dfrac{k_{ky}}{M_y}y - \dfrac{k_{by}}{M_y}\dot{y} + \dfrac{c_y k_{iy}}{M_y}u_y - \dfrac{L_y}{M_y}, \end{cases} \tag{2}$$

where $M_x = m_b + m_p + m_l$ and $M_y = m_p + m_l$; L_x and L_y are the lumped uncertainties regarded as follows:

$$\begin{cases} L_x = \Delta k_{kx} x + \Delta k_{bx} \dot{x} - \Delta c_x k_{ix} u_x + F_{fx} + F_{dx}, \\ L_y = \Delta k_{ky} y + \Delta k_{by} \dot{y} - \Delta c_y k_{iy} u_y + F_{fy} + F_{dy}. \end{cases} \tag{3}$$

In equation (1), the practical control characteristics of the VCMs are nonlinear because the system coefficients described above may vary due to the changes in operating temperature and duration, though the VCMs-driven dual-axis positioning stage can be presented with a state-space model. Moreover, the lumped uncertainties L_x and L_y cannot be measured exactly. Therefore, designing a model-free control method is important to control the VCMs-driven dual-axis positioning stage with a stable and precise nonlinear contour tracking performance for the practical applications.

3. CSA with an Adaptive Strategy

CSA is a metaheuristic evolutionary algorithm based on the aggressive reproduction of a cuckoo species with a Lévy flight behavior. Three idealized characteristic rules are assumed as follows to formulate the CSA [13–16]:

(a) Each cuckoo bird lays one egg in a randomly selected host nest, representing a solution to the optimization problem.

(b) Some of these nests contain high-quality eggs, representing good solutions, which are preserved for the next generation.

(c) The number of available host nests is fixed in the ecosystem, and the probability of alien eggs discovered by the host bird is $P_a \in [0, 1]$. When the host bird finds the alien eggs, it destroys the egg or abandons the old nest and builds a new one in another place.

3.1. Principle of CSA. From the optimization perspective, cuckoo birds correspond to the particles assigned to find solutions, and cuckoo eggs indicate the candidate solutions for an optimization problem. In the CSA, the random step of cuckoo birds is characterized by a Lévy flight, indicating that the step length of the flight behavior follows the Lévy distribution; consequently, the CSA realizes a "random walk" and a "long jump" among their flights [15]. In this regard, the CSA can avoid obtaining an unreliable local optimal solution and shorten the convergence time required to reach a global optimal solution.

An unconstrained optimization problem can be stated as follows:

$$\text{find } \mathbf{x} = [x_1, x_2, \ldots, x_D], \quad \text{which maximizes } J(\mathbf{x}), \tag{4}$$

where $\mathbf{x}$ is the individual nest position, D is the optimized variable dimension, and J is an objective function. In the CSA, the update of the egg position is given according to a Lévy flight as follows [13–16]:

$$\mathbf{x}_{i,k+1} = \mathbf{x}_{i,k} + \alpha \oplus \text{Lévy}(\beta), \tag{5}$$

where $i = 1, 2, \ldots, N_p$ is the population size, k is the current index for the generation iteration, $\oplus$ is entry-wise multiplication, $\alpha > 0$ is a step size related to the scales of the problem of interest, and $1 \le \beta \le 3$ is a parameter used to formulate the Lévy distribution and it is considered to be 1.5 in this study. Then, the step length ς is defined as

$$\varsigma = \frac{\mu}{|\nu|^{(1/\beta)}}, \tag{6}$$

where μ and ν are random numbers derived from normal distribution as

$$\begin{aligned} \mu &\sim N\left(0, \sigma_\mu^2\right), \\ \nu &\sim N\left(0, \sigma_\nu^2\right), \end{aligned} \tag{7}$$

$$\begin{aligned} \sigma_\mu &= \left\{ \frac{\Gamma(1+\beta) \times \sin(\pi\beta/2)}{\Gamma[(1+\beta)/2] \times \beta \times 2^{(\beta-1)/2}} \right\}^{(1/\beta)}, \\ \sigma_\nu &= 1. \end{aligned} \tag{8}$$

where σ_μ is derived by using Mantegna's algorithm for symmetric distributions and $\Gamma(\cdot)$ is a Gamma function. Then, the step size $\mathbf{s}$ is calculated as

$$\mathbf{s}_{i,k} = \alpha \cdot \varsigma \cdot \left(\mathbf{x}_{i,k} - \mathbf{x}_b\right), \tag{9}$$

where $\mathbf{x}_b$ is the current best solution. Thus, the update of the egg position as shown in equation (5) can be formulated:

$$\mathbf{x}_{i,k+1} = \mathbf{x}_{i,k} + r \cdot \mathbf{s}_{i,k}, \tag{10}$$

where r is a random value following the normal distribution $N(0, 1)$. Figure 3 shows the typical trajectory of a three-dimension random Lévy flight path by using equations (5)–(10). Afterward, the fitness values of $J(\mathbf{x}_{i,k+1})$ and $J(\mathbf{x}_{i,k})$ are compared. If $J(\mathbf{x}_{i,k+1}) > J(\mathbf{x}_{i,k})$ holds, the ith solution is replaced, and the new solution is accepted as $\mathbf{x}_{i,k+1}$. In addition, the parameter P_θ is set as the threshold of discovery probability that the cuckoo's eggs are found by a host bird. The host bird builds nests at new locations according to

$$\mathbf{x}_{i,k+1} = \begin{cases} \mathbf{x}_{i,k} + r \cdot \left(\mathbf{x}_{q,k} - \mathbf{x}_{j,k}\right), & \text{if } P > P_\theta, \\ \mathbf{x}_{i,k}, & \text{else}, \end{cases} \tag{11}$$

where $\mathbf{x}_{q,k}$ and $\mathbf{x}_{j,k}$ are two randomly selected different solutions in the kth iteration and P is a uniform random number distributed in [0, 1]. Similarly, if the fitness value of the new solution is better than the old one, then the new solution $\mathbf{x}_{i,k+1}$ is used to replace the old one $\mathbf{x}_{i,k}$.

3.2. Adaptive Strategy of the CSA. An adaptive strategy based on Rechenberg's 1/5 criteria is utilized to enhance the evolution and adaptation efficiency of the CSA [14]. With the adaptive strategy, step size and discovery probability are dynamically tuned during optimization. First, the improvement rate ζ is defined as follows:

$$\zeta = \frac{N_r}{N_p}, \tag{12}$$

where N_r is the number of all cuckoo birds whose fitness values are improved after evolution. Thus, the step size α and discovery probability P_θ can be further updated as

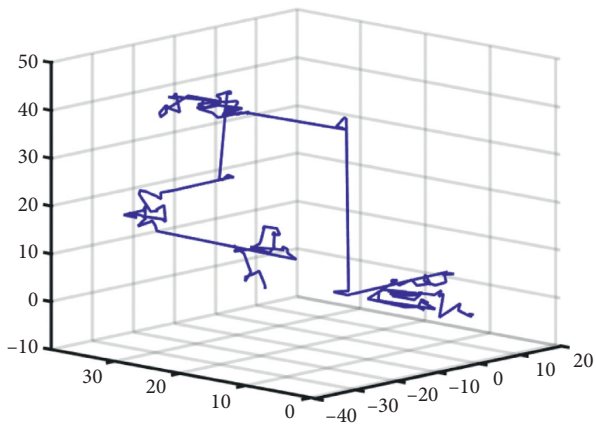


Figure 3: Typical trajectory of a three-dimension random Lévy flight path.

$$\alpha_{k+1} = \begin{cases} \alpha_k \times f_\alpha, & \zeta > \alpha_u, \\ \alpha_k, & \alpha_l \le \zeta \le \alpha_u, \\ \dfrac{\alpha_k}{f_\alpha}, & \zeta < \alpha_l, \end{cases} \tag{13}$$

$$P_{\theta,k+1} = \begin{cases} P_{\theta,k} \times f_P, & \zeta > P_u, \\ P_{\theta,k}, & P_l \le \zeta \le P_u, \\ \dfrac{P_{\theta,k}}{f_P}, & \zeta < P_l, \end{cases} \tag{14}$$

where α_u and α_l are the upper and lower thresholds of ζ with respect to α; P_u and P_l are the upper and lower thresholds of ζ with respect to the discovery probability; and $1 \le f_\alpha \le 2$ and $1 \le f_P \le 2$ are the learning factors of α and P_θ, respectively.

According to equations (13) and (14), α and P_θ are increased to strengthen the global exploration ability when ζ is large. This result indicates that the current solution space is relatively monotonous and smooth. On the contrary, α and P_θ are decreased to enhance the local exploitation ability when ζ is small. This result suggests that the optimal solution may be in the surrounding search area near the current solution. In this regard, the local exploitation and global exploration abilities of the CSA can be well balanced to deal with the diversification and intensification of a population, thereby avoiding the premature convergence.

4. Control System Designs of the VCMs-Driven Dual-Axis Positioning Stage

First, typical PID and FFPID control strategies are adopted in this study to control the VCMs-driven dual-axis positioning stage for nonlinear contour tracking. Subsequently, a VOFFPID is proposed to improve the stability and accuracy of the contour tracking performance under system uncertainties, including parameter variations, cross-coupled interferences, and friction forces [27]. With the help of online-tuned control parameters, the system uncertainties can be compensated, and the high-precision nonlinear contour tracking performance can be guaranteed.

4.1. Typical PID Control. The popularity of IO PID (IOPID) controllers as expressed in equation (15) can be attributed partly to their favorable performance in a wide range of applicability and partly to their functional simplicity, which allows engineers to operate them in an easy and straightforward manner. As for PID controller, the proportional (P) action amplifies errors, the integral (I) action accumulates errors, and the differential (D) action calculates the change in errors. In this study, the PID controller compares the actual mover positions x and y with the reference contour positions x_d and y_d to obtain the error signals e_x and e_y. After that, it accumulates the results of the P, I, and D actions as below [7]:

$$u_j(t) = K_{\mathrm{P}j} e_j(t) + K_{\mathrm{I}j} \int_0^t e_j(\tau)\mathrm{d}\tau + K_{\mathrm{D}j} \frac{\mathrm{d}}{\mathrm{d}t} e_j(t), \tag{15}$$

where t denotes the current time; $j = x, y$ represent the X-axis and Y-axis of the VCMs-driven dual-axis positioning stage, respectively; u_x and u_y denote the control signals; $K_{\mathrm{P}j}$, $K_{\mathrm{I}j}$, and $K_{\mathrm{D}j}$ denote the P, I, and D control parameters, respectively; and e_x and e_y indicate the tracking errors defined as $e_x = x_d - x$ and $e_y = y_d - y$, respectively.

4.2. FO Integral and Differential Definitions. FO integral and differential operators are defined in the following [7, 18, 25]:

$${}_aD_t^\lambda = \begin{cases} \dfrac{\mathrm{d}^\lambda}{\mathrm{d}t^\lambda}, & \lambda > 0, \\ 1, & \lambda = 0, \\ \displaystyle\int_a^t (\mathrm{d}\tau)^{-\lambda}, & \lambda < 0, \end{cases} \tag{16}$$

in which D is the fractional calculus operator; λ is the fractional order; and a and t represent the operation range. The three major FO integral and differential definitions are the Caputo, Grunwald-Letnikov (GL), and Riemann–Liouville (RL) definitions. The operator given in equation (16) applied to the $f(t)$ function leads to an extended Caputo form, which can be derived as follows [28, 29]:
tcusts6 50

$${}_aD_t^\lambda f(t) = \frac{1}{\Gamma(m-\lambda)} \int_a^t \frac{f^{(m)}(\tau)}{(t-\tau)^{\lambda+1-m}} \mathrm{d}\tau, \quad m-1 \le \lambda < m, \tag{17}$$

where m is an integer such that $m > \lambda$. Moreover, the λth-order RL FO integral of $f(t)$ is defined as follows [30]:

$${}_aD_t^{-\lambda} f(t) = \frac{1}{\Gamma(\lambda)} \int_a^t (t-\tau)^{\lambda-1} f(\tau)\mathrm{d}\tau. \tag{18}$$

Similarly, the RL FO differential of $f(t)$ is defined as

$${}_aD_t^\lambda f(t) = \frac{1}{\Gamma(m-\lambda)} \frac{\mathrm{d}^m}{\mathrm{d}t^m} \int_a^t \frac{f(\tau)}{(t-\tau)^{\lambda+1-m}} \mathrm{d}\tau. \tag{19}$$

By contrast, the λth-order GL FO operation based on finite differences is defined as follows [29]:

$${}_aD_t^\lambda f(t) = \lim_{h \longrightarrow 0} h^{-h} \sum_{j=0}^{[t-a/h]} (-1)^j \binom{\lambda}{j} f(t-jh), \tag{20}$$

where [.] is the integer part, h is the time increment, and () is the fractional binomial coefficient defined as

$$\binom{\lambda}{j} = \frac{\Gamma(\lambda+1)}{\Gamma(j+1)\cdot\Gamma(\lambda-j+1)}. \tag{21}$$

Intuitively, integral and differential operations with fractional orders can provide a higher degree of freedom to the control parameters than those with integer orders. As a result, the control performance of PID control system can be enhanced by properly selecting fractional integral and differential orders. For convenience, the FO operator ${}_aD_t^{\lambda}$ is noted as D^{λ} in the subsequent sections.

4.3. Developed FFPID Control System. In the case of a nonlinear and disturbed system, the conventional IOPID control strategy is difficult to concurrently obtain a high control performance level and maintain good robustness because of its linear structure [7]. To improve the control performances, smoothness and robustness of the PID control system, the FFPID control, which combines the merits of PID control, FO operations, and FLS, is adopted and illustrated in Figure 4 in this study. In Figure 4, a_j and b_j are the fractional differential and integral orders, respectively; $K_{\mathrm{P}j}$ and $K_{\mathrm{D}j}$ can be considered the input scaling factors; and $K_{\mathrm{I}j}$ can be regarded as the output scaling factor. The inputs of the FLS are the tracking error e_j multiplied by $K_{\mathrm{P}j}$ and the fractional differential of the tracking error $D^{aj}e_j$ multiplied by $K_{\mathrm{D}j}$, which can be regarded as a FO proportional-differential (FOPD) controller. The relationship between the inputs and output of the FLS is specified with the table of the fuzzy rules as given in Table 2 in which the fuzzy linguistic values NL, NM, NS, ZO, PS, PM, and PL indicate negatively large, negatively medium, negatively small, zero, positively small, positively medium, and positively large, respectively [7]. Figure 5 illustrates the membership functions for the inputs and output of FLS in which the horizontal range was designed on the basis of the prior experimental tests to effectively cover the input and output signals [7]. In this study, the triangular membership functions, which can be easily configured with regard to the linear shape and fewer parameters, were selected to ease the computational burden and speed up the control process. Thus, the output of the FLS $u_{\mathrm{FPD}j}$ can be derived according to the designed fuzzy rules with the center of gravity defuzzification method as follows:

$$u_{\mathrm{FPD}j} = \frac{\sum_{k=1}^{n}\mu_c(\sigma_k)\sigma_k}{\sum_{k=1}^{n}\sigma_k}, \tag{22}$$

where c indicates a logical union set of the conclusion fuzzy sets of the fired fuzzy rules; σ_k is a value between the minimum and maximum values of the abscissa of c defined on the universe of discourse; $\mu_c(\sigma_k)$ is the firing strength of c for the point σ_k; and n is the number of the samples.

The final control signal of the FFPID control system u_j is the sum of the output of FLS $u_{\mathrm{FPD}j}$ multiplied by λ_j and the fractional integral of the output of FLS $u_{\mathrm{FPD}j}$ multiplied by $K_{\mathrm{I}j}$:

$$u_j(t) = \lambda_j u_{\mathrm{FPD}j}(t) + K_{\mathrm{I}j}D^{-bj}u_{\mathrm{FPD}j}(t). \tag{23}$$

In equations (22) and (23), the whole FFPID controller can be considered a combination of the fuzzy FOPD controller $u_{\mathrm{FPD}j}$ in the first half and the FO proportional-integral (FOPI) controller u_j in the second half. The benefits of the FFPID controller are adjustability and flexibility when these two controllers are combined. On the other hand, as seen from Figure 4, the integral operator $D^{-bj}e_j$ can be regarded a low-pass filter of the error signal e_j. When b_j is appropriately selected, the steady-state error can be suppressed effectively [18]. Besides, the differential operator $D^{aj}e_j$ can be regarded a high-pass filter of e_j. A proper a_j can accelerate the dynamic response of the VCMs-driven dual-axis positioning system. Therefore, the contour tracking responses with a conventional IOPID controller can be enhanced by adding the well-defined fractional orders a_j and b_j regarding the specified dynamics of the VCMs-driven dual-axis positioning stage.

4.4. Proposed VOFFPID Control System. The control gains (i.e., $K_{\mathrm{P}j}$, K_{Ij}, and K_{Dj}), along with fractional orders of differentiation (i.e., a_j) and integration (i.e., b_j), are tuned to obtain the optimum contour tracking performance of the VCMs-driven dual-axis positioning system. Hence, a VOFFPID controller is further proposed, in which the control parameters $\{K_{\mathrm{P}j}, K_{\mathrm{I}j}, K_{\mathrm{D}j}, a_j, b_j\}$ are dynamically tuned with the CSA with an adaptive strategy. In the CSA application, the most crucial step is to choose the objective function for evaluating the fitness value of each host nest. In this study, an absolute tracking error is employed to design the objective function. Thus, the optimization problem arising in this study can be expressed by rewriting equation (4) as follows:

$$\text{Find } \mathbf{x} = \left[K_{\mathrm{P}j}, K_{\mathrm{I}j}, K_{\mathrm{D}j}, a_j, b_j\right], \quad \text{which maximizes } J(\mathbf{x}) = \frac{1}{\varepsilon + \left|e_c(\mathbf{x})\right|}, \tag{24}$$

where ε is a small positive constant and e_c is a contour tracking error defined as follows:

$$e_c(\mathbf{x}) = \sqrt{e_x(\mathbf{x})^2 + e_y(\mathbf{x})^2}. \tag{25}$$

According to the design of the object function shown in equation (24), $K_{\mathrm{P}j}$, K_{Ij}, K_{Dj}, a_j, and b_j can be updated dynamically to minimize the contour tracking error e_c via the CSA.

In the beginning of the VOFFPID control system, several nest positions $\mathbf{x}$ are selected randomly within the specific searching ranges. Then, each vector $\mathbf{x}$ is sequentially applied to the VOFFPID controller, and the corresponding tracking performance is evaluated via the object function J. Lastly, the vector with the highest fitness value is selected for the VCMs-driven dual-axis positioning system. As a result, the VOFFPID controller can achieve favorable robustness against uncertainties and external disturbances.

5. Experimental Results

5.1. Experimental Setup. Figure 6 shows the experimental setup of the VCMs-driven dual-axis positioning system, which consists of a newly developed dual-axis positioning stage, power supplies, servo drivers (Elmo Cello 5/60), and a

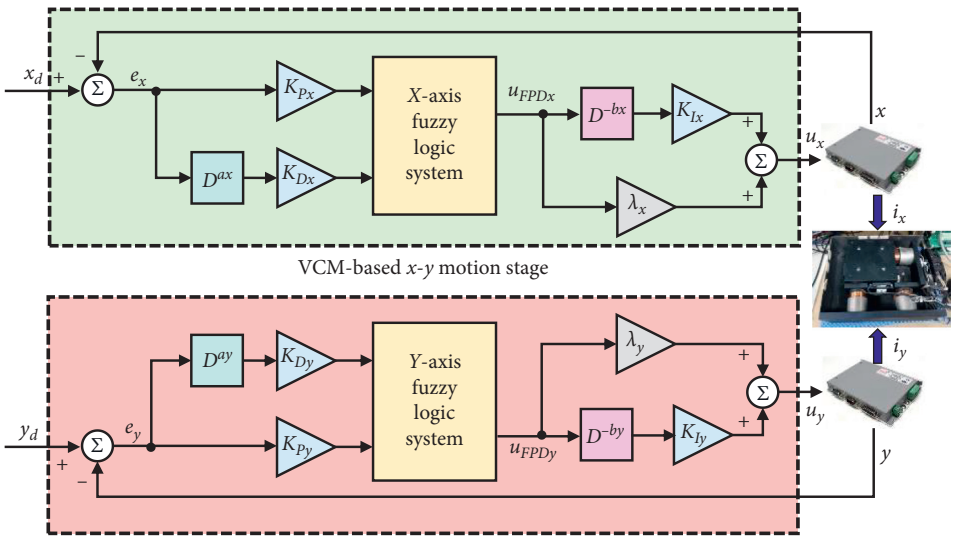


Figure 4: Control diagram of the VCMs-driven dual-axis positioning control system using FFPID controller.

Table 2: Fuzzy rule table.

$K_{Dj}D^{aj}e_j$ \ $K_{pj}e_j$	NL	NM	NS	ZR	PS	PM	PL
PL	ZR	PS	PM	PL	PL	PL	PL
PM	NS	ZR	PS	PM	PL	PL	PL
PS	NM	NS	ZR	PS	PM	PL	PL
ZR	NL	NM	NS	ZR	PS	PM	PL
NS	NL	NL	NM	NS	ZR	PS	PM
NM	NL	NL	NL	NM	NS	ZR	PS
NL	NL	NL	NL	NL	NM	NS	ZR

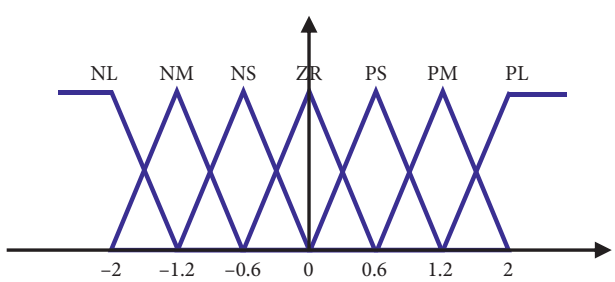


Figure 5: Membership functions of the input and output variables of FLS.

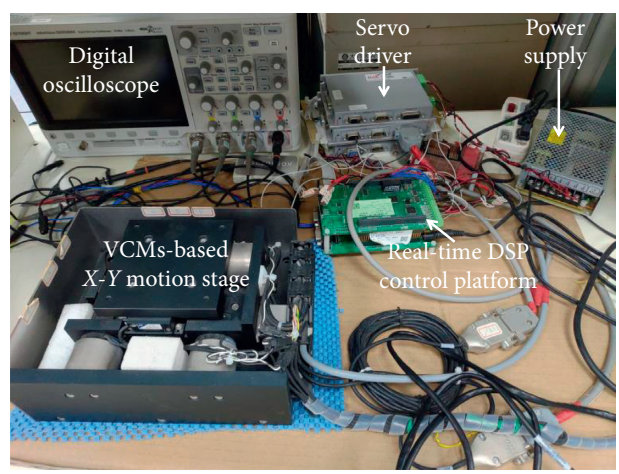


Figure 6: Experiment setup of the VCMs-driven dual-axis positioning stage.

TMS320F28377 digital signal processor (DSP; Texas Instruments) [7]. The real-time control software developed in the DSP comprises one main program and one interrupt service routine (ISR). In the main program, parameters and I/O initializations are initially established, and the interrupt interval for the ISR is set. When the interrupt is enabled, the ISR, with 1 ms execution frequency, calculates the mover position from the encoder interfaces and then determines the control signals through the designed PID, FFPID, and VOFFPID control systems. After that, the control signals are sent to the servo drivers via the 14-bit resolution digital-to-analog converters (DACs) of the DSP. Afterward, the servo drivers convert the control signals to drive currents so that the VCMs can produce the required thrust force for high-precision contour tracking. In this study, a flower contour and a window contour are designed for the reference nonlinear contour commands as shown in Figures 7 and 8, respectively.

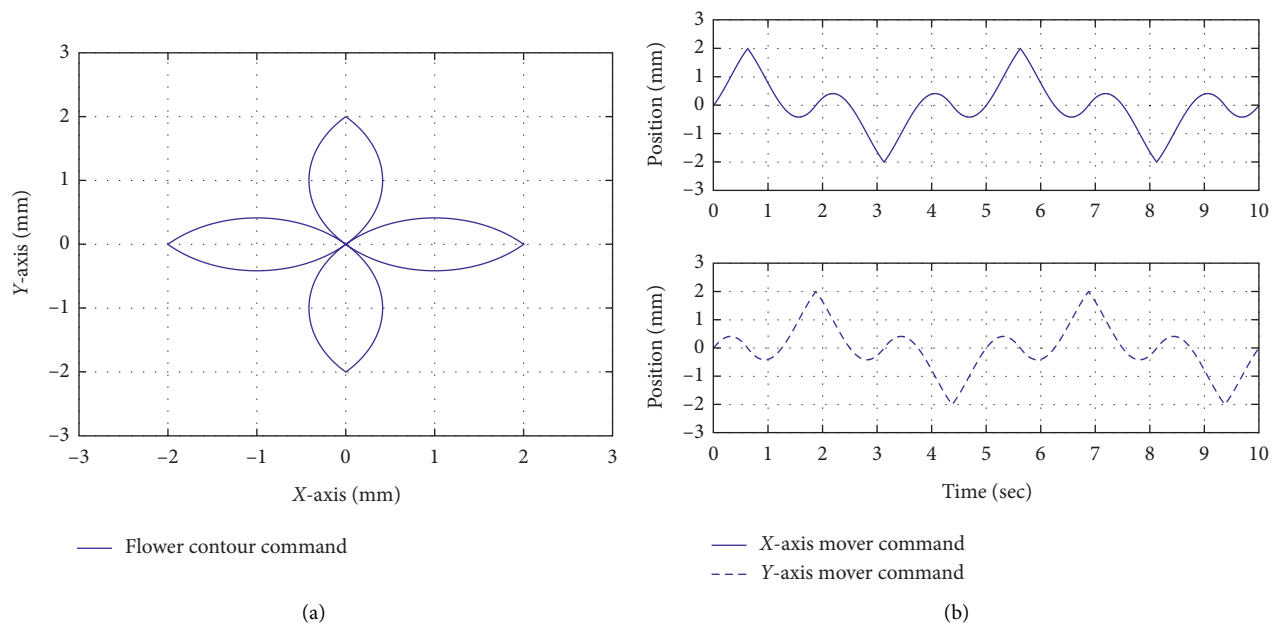


Figure 7: Design of flower contour command. (a) Flower contour in X-Y-axes; (b) mover commands of flower contour in X-axis and Y-axis.

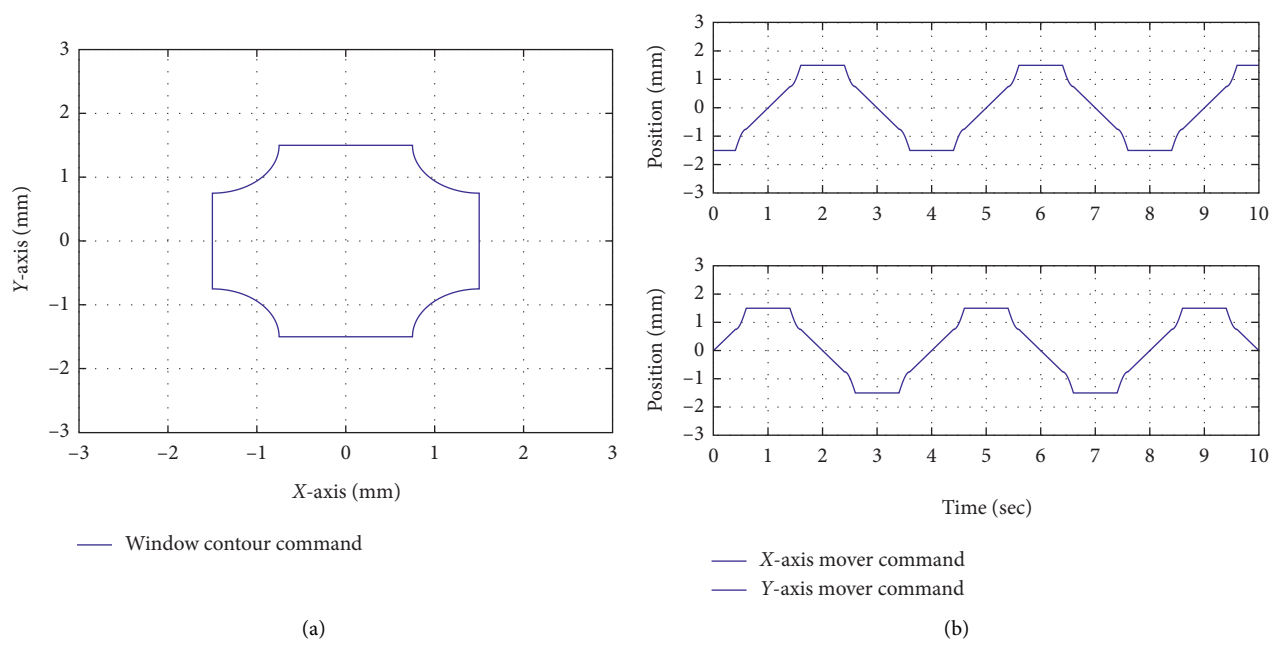


Figure 8: Design of window contour command. (a) Window contour in X-Y-axes; (b) mover commands of window contour in X-axis and Y-axis.

Table 3: Formulas $A_n(z^{-1}, \lambda)$ for $n = 0$, 1, 3, and 5.

n	$A_n(z^{-1}, \lambda)$
0	1
1	$-\lambda z^{-1} + 1$
3	$-(1/3)\lambda z^{-3} + (1/3)\lambda^2 z^{-2} - \lambda z^{-1} + 1$
5	$-(1/5)\lambda z^{-5} + (1/5)\lambda^2 z^{-4} - ((1/3)\lambda + (1/15)\lambda^3) z^{-3} + (2/5)\lambda^2 z^{-2} - \lambda z^{-1} + 1$

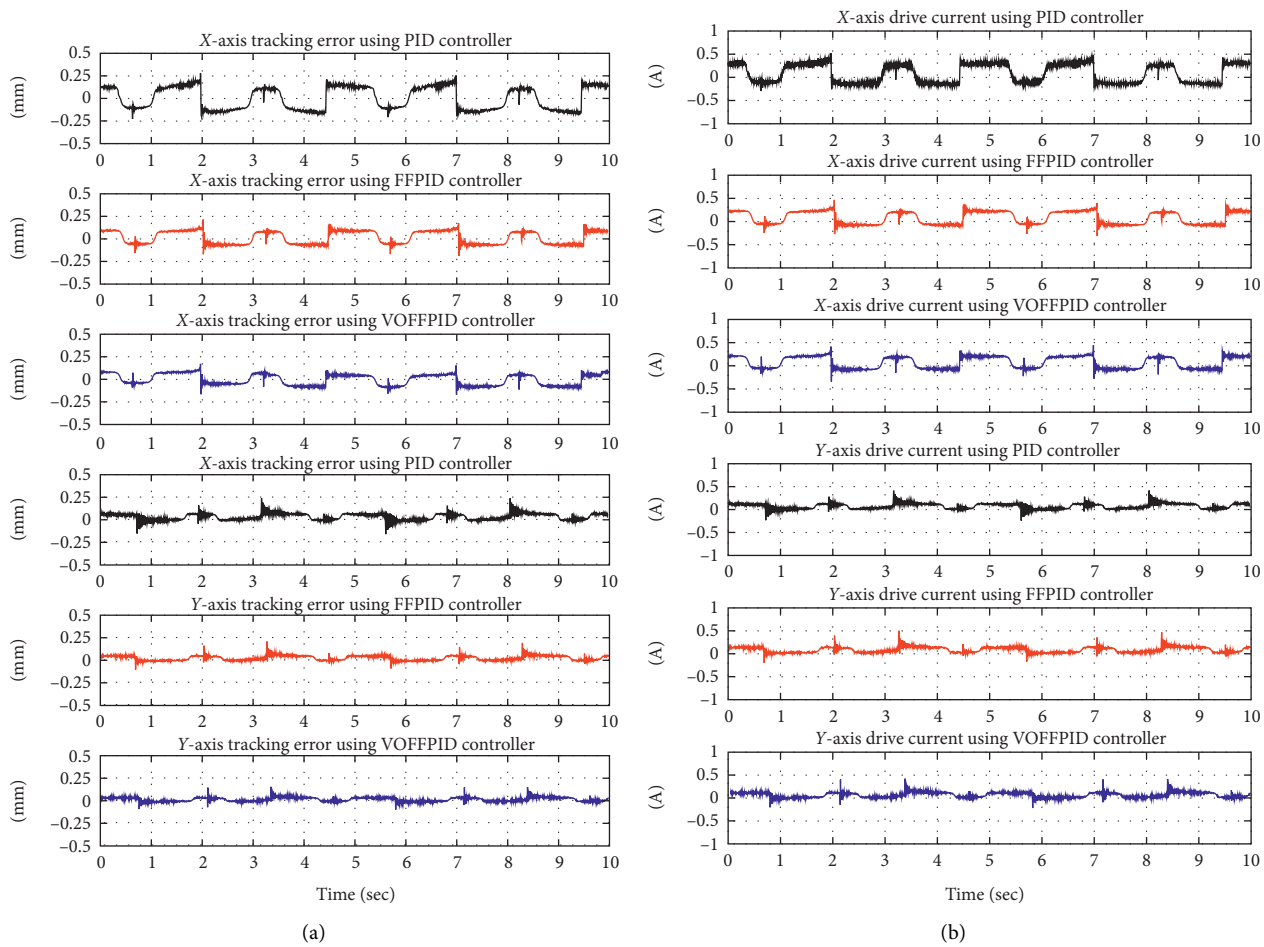


Figure 9: Flower contour tracking results of the VCMs-driven dual-axis positioning stage using PID, FFPID, and VOFFPID controllers in Case 1. (a) Tracking errors; (b) drive currents.

The maximum, average, and standard deviation of the contour tracking error T_m, T_A, and T_S are measured as follows to compare the different positioning performance levels of the PID, FFPID, and VOFFPID control systems [7]:

$$T_M = \max_I e_c(I),$$
$$T_A = \sum_{I=1}^{N_T} \frac{e_c(I)}{N_T}, \tag{26}$$
$$T_S = \sqrt{\sum_{I=1}^{N_T} \frac{[e_c(I) - T_A]^2}{N_T}},$$

where I is the current iteration number and N_T is the total number of iterations. Moreover, two conditions are tested in this study: nominal (Case 1) and payload (Case 2) cases. In Case 2, one payload with a 5 kg weight is added to the mover.

5.2. Discretization of FO Integral and Differential. As seen from the FO definitions shown in (17)–(20), the Laplace transform of the FO differential and integral of function $f(t)$ can be represented by $s^\lambda F(s)$, where $s = j\omega$ is the Laplace transform operator. Hence, the Tustin method is used to obtain the coefficients and the form of the direct discretization of s^λ. To simplify the presentation, only the recursive formula for a positive λ is considered. Thus, the continuous Laplace operator can be replaced by a generating function as follows [18, 25]:

$$s^\lambda = \left(\omega\left(z^{-1}\right)\right)^\lambda = \left(\frac{2}{T}\right)^\lambda \left(\frac{1 - z^{-1}}{1 + z^{-1}}\right)^\lambda = \left(\frac{2}{T}\right)^\lambda \lim_{n\to\infty} \frac{A_n\left(z^{-1}, \lambda\right)}{A_n\left(z^{-1}, -\lambda\right)}, \tag{27}$$

where z is the shifting operator and T is the sampling period:

$$A_o\left(z^{-1}, \lambda\right) = 1,$$
$$A_n\left(z^{-1}, \lambda\right) = A_{n-1}\left(z^{-1}, \lambda\right) - c_n z^n A_{n-1}(z, \lambda), \tag{28}$$
$$c_n = \begin{cases} \dfrac{\lambda}{n}, & n \text{ is odd;} \\ 0, & n \text{ is even.} \end{cases}$$

Consequently, the Laplace operator can be approximated to derive the FO integral and differential based on any given order of approximation n, as follows:

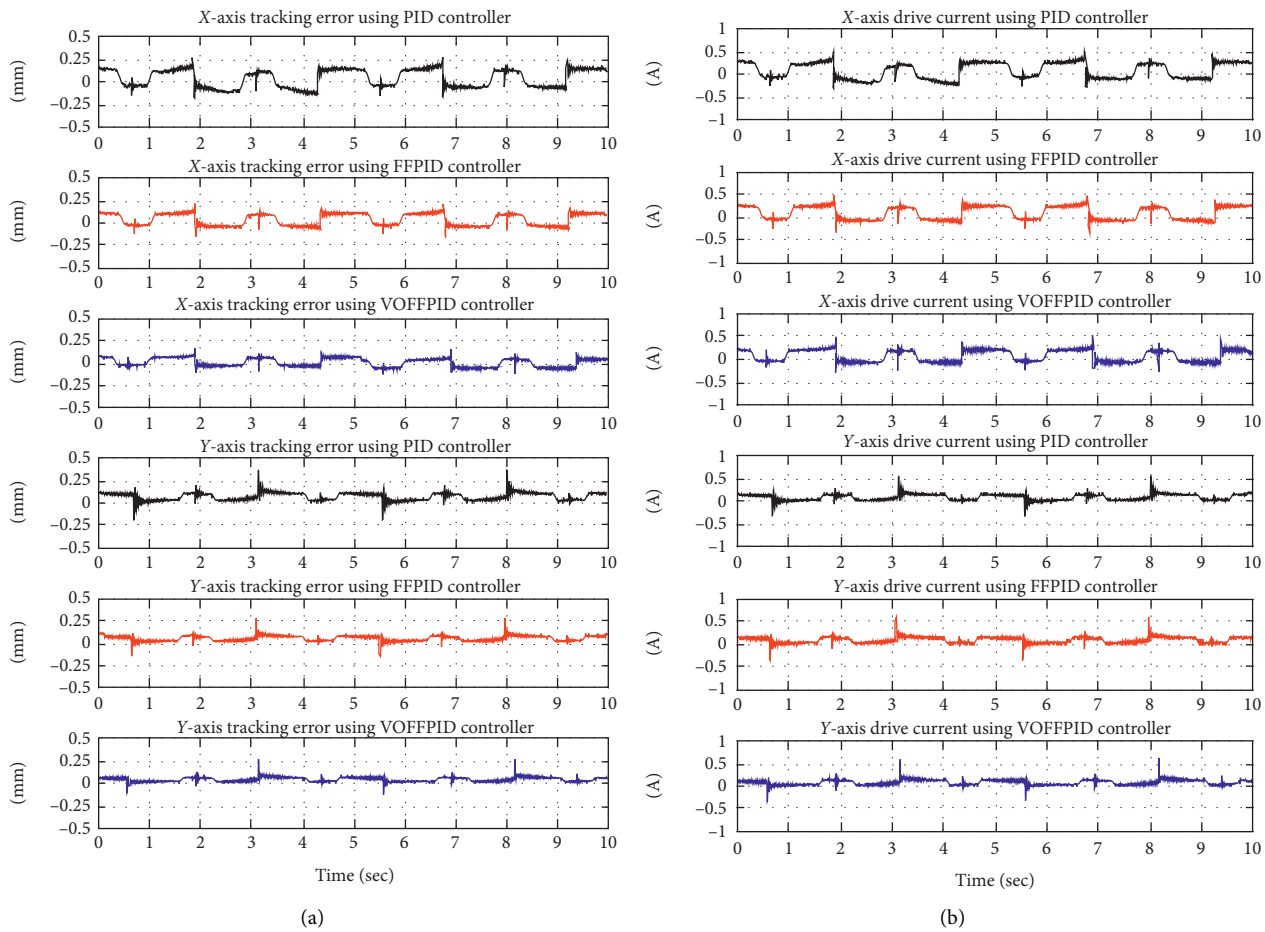


Figure 10: Flower contour tracking results of the VCMs-driven dual-axis positioning stage using PID, FFPID, and VOFFPID controllers in Case 2. (a) Tracking errors; (b) drive currents.

$$s^{\lambda} \approx \left(\frac{2}{T}\right)^{\lambda} \frac{A_n\left(z^{-1},\lambda\right)}{A_n\left(z^{-1},-\lambda\right)}. \tag{29}$$

Thus, the FO operations can be realized via the digital implementation. Table 3 lists the expressions of $A_n(z^{-1},\lambda)$ for $n = 0$, 1, 3, and 5.

5.3. Experimental Results. In the experiment, the control parameters of the PID controller were chosen as $K_{Px} = 25$, $K_{Ix} = 110$, $K_{Dx} = 10$, $K_{Py} = 25$, $K_{Iy} = 110$, and $K_{Dy} = 12$, respectively. Moreover, the control parameters of the FFPID controller are selected as $K_{Px} = 25$, $K_{Ix} = 100$, $K_{Dx} = 9$, $\lambda_x = 1$, $a_x = 0.6$, $b_x = 0.5$, $K_{Py} = 25$, $K_{Iy} = 100$, $K_{Dy} = 11$, $\lambda_y = 1$, $a_y = 0.5$, and $b_y = 0.5$, respectively. In addition, a third-order approximation was used for the FO digital realization; that is, $n = 3$. In this study, the control parameters were selected on the basis of several trials to achieve the favorable transient responses, considering the requirement of steady-state stability. However, designing an optimal set for all the control parameters is difficult because of the occurrence of uncertainties. Additionally, the PID and FFPID controllers cannot maintain ideal positioning performances by adopting the constant control parameters.

5.3.1. Flower Contour Tracking Results. The experimental results, including the tracking errors and drive currents of the VCMs-driven dual-axis positioning stage controlled by the PID, FFPID, and VOFFPID control systems due to the flower contour tracking in Cases 1 and 2, are shown in Figures 9 and 10, respectively. As can be seen from Figures 9(a) and 10(a), the mover of the stage can be successfully controlled by all the controllers to track the reference nonlinear contour shown in Figure 7. Furthermore, the drive currents in Case 2 are larger than those in Case 1, so a higher thrust force for the additional payload can be generated. The maximum tracking errors obtained in Case 1 for the PID, FFPID, and VOFFPID control systems were 0.2807 mm, 0.2363 mm, and 0.1752 mm, respectiely, whereas those obtained in Case 2 were 0.3973 mm, 0.2986 mm, and 0.2731 mm, respectively.

The tracking errors of the PID control system were unfavorable because of the large tracking errors. Although selecting larger control gains can diminish the amplutide of tracking errors, the excessive aggressive control gains may result in the oscillation of control responses.

As seen in Figures 9 and 10, the FFPID with two well-designed variables a and b and FLS can derive more effective and smooth control signals to restrain the contouring errors

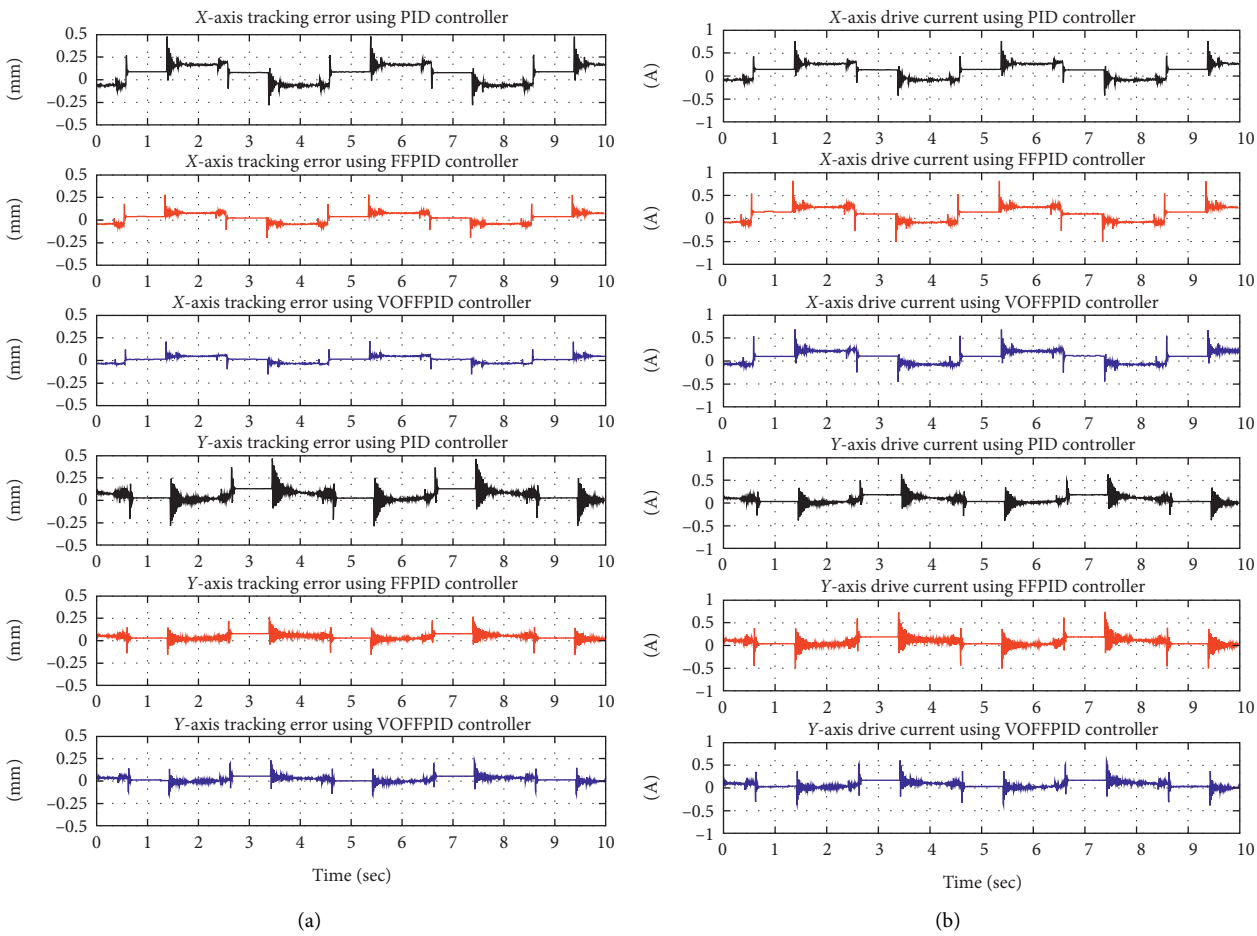


Figure 11: Window contour tracking results of the VCMs-driven dual-axis positioning stage using PID, FFPID, and VOFPID controllers in Case 1. (a) Tracking errors; (b) drive currents.

related to the specified dynamics of VCMs and the possible occurrence of uncertainties during the flower contour tracking. The corresponding tracking errors were reduced compared with those of the PID controller. Moreover, the FFPID controller has a good ability to diminish the effect of the disturbance, as shown in Figures 9(a) and 10(a). Although the control parameters of the FFPID controller were selected with several trials, the maximum and average tracking errors in the nominal and payload conditions are obviously reduced by the self-tuned strategy.

5.3.2. Window Contour Tracking Results. The experimental results due to the window contour tracking in Cases 1 and 2 are shown in Figures 11 and 12, respectively. Similar behaviors on the tracking responses of flower contour tracking can be observed. As seen from Figures 11(a) and 12(a), the maximum tracking errors obtained in Case 1 for the PID, FFPID, and VOFFPID control systems were 0.4877 mm, 0.2843 mm, and 0.2344 mm, respectively, whereas those obtained in Case 2 were 0.5944 mm, 0.3512 mm, and 0.3098 mm, respectively. From the comparison in Figures 11(a) and 12(a), the tracking performances of the PID are evidently deteriorated when the contour command changes instantaneously. In contrast, the proposed VOFFPID demonstrates its robustness in the tracking performance during both test conditions. On the other hand, the control oscillations in the PID control system as shown in Figures 11(b) and 12(b) are evident due to its inefficient tracking ability. As opposed to the PID, more effective and smooth control signal was derived by the proposed VOFFPID to carry out the best control performance.

The experimental results and observations reveal that the optimized control parameters can improve the tracking performance in practical control applications. In Figures 9–12, the best control performance of the VOFFPID controller due to the flower and window contours under the nominal and payload conditions can be clearly observed. The improvement of the proposed VOFFPID controller in terms of the contour tracking accuracy is significant compared with that given by traditional PID and FFPID controllers.

The contour tracking performance measures of the PID, FFPID, and VOFFPID control systems for the tracking of the flower and window reference nonlinear contours are shown in Tables 4 and 5, respectively. They indicate that the FFPID controller with the integration of the PID control, FO operation, and FLS outperforms the conventional PID controller. Moreover, the proposed VOFFPID controller further

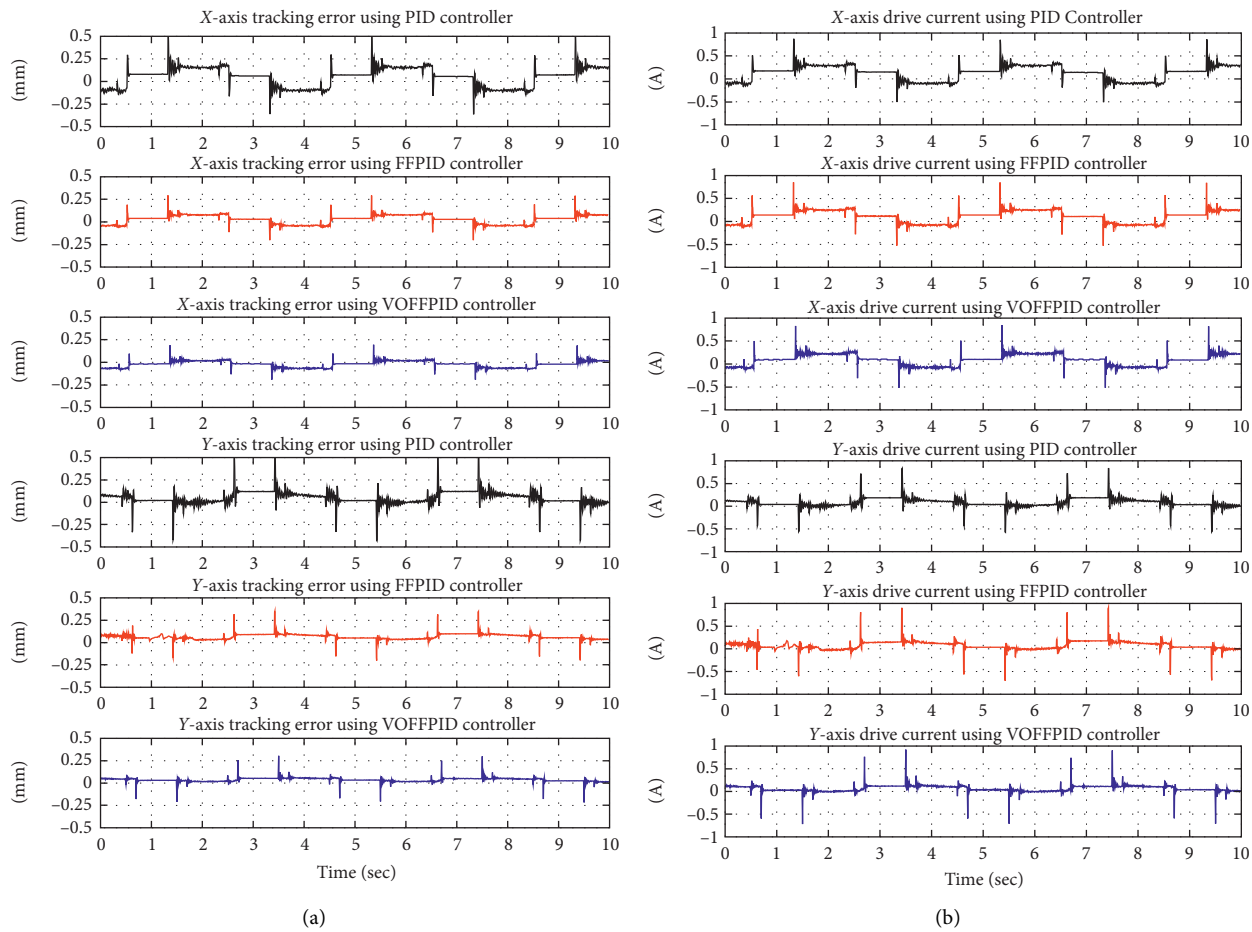


Figure 12: Window contour tracking results of the VCMs-driven dual-axis positioning stage using PID, FFPID, and VOFFPID controllers in Case 2. (a) Tracking errors; (b) drive currents.

Table 4: Contour tracking performance measures of Case 1.

Controllers	Commands					
	Flower contour (mm)			Window contour (mm)		
	T_M	T_A	T_S	T_M	T_A	T_S
PID	0.2807	0.1316	0.0336	0.4877	0.1389	0.0555
FFPID	0.2363	0.0791	0.0233	0.2843	0.0756	0.0263
VOFFPID	0.1752	0.0649	0.0210	0.2344	0.0471	0.0249

Table 5: Contour tracking performance measures of Case 2.

Controllers	Commands					
	Flower contour (mm)			Window contour (mm)		
	T_M	T_A	T_S	T_M	T_A	T_S
PID	0.3973	0.1312	0.0490	0.5944	0.1355	0.0597
FFPID	0.2986	0.0921	0.0383	0.3512	0.0908	0.0253
VOFFPID	0.2731	0.0682	0.0233	0.3098	0.0535	0.0289

improves the tracking performance of the FFPID controller because all the control parameters were globally and dynamically optimized by the CSA algorithm. The VOFFPID controller apparently exhibits a high-precision contour tracking performance by effectively handling the payload and uncertainty during control processes.

6. Conclusions

In this study, a VOFFPID control system is successfully developed and applied to control the mover position of a new VCMs-driven dual-axis positioning stage for tracking nonlinear reference contours. First, the structural and operating principles of the stage are introduced. Then, the CSA with the adaptive strategy for the optimization of control parameters is described. Subsequently, the theoretical bases of the PID, FFPID, and VOFFPID control systems are given in detail. With an additional degree of freedom to the control parameters and FLS operation, the FFPID controller can upgrade the contouring performances of the PID controller. Moreover, in the proposed VOFFPID controller, the CSA with the adaptive strategy can enhance the robustness of the FFPID controller by tuning the control parameters online. The experimental results subjected to different performance measures are given to verify the effectiveness of the proposed VOFFPID controller.

Data Availability

The experimental data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors acknowledge the financial support of the Ministry of Science and Technology in Taiwan, R.O.C., under Grant no. MOST 108-2221-E-003-022-MY2.

References

[1] S. Das, I. Pan, S. Das, and A. Gupta, "A novel fractional order fuzzy PID controller and its optimal time domain tuning based on integral performance indices," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 2, pp. 430–442, 2012.

[2] R. Sharma, P. Gaur, and A. P. Mittal, "Performance evaluation of cuckoo search algorithm based FOPID controllers applied to a robotic manipulator with actuator," in *Proceedings of the 2015 International Conference on Advances in Computer Engineering and Applications*, pp. 356–363, Ghaziabad, India, March 2015.

[3] Q. Sun, Z. Qi, J. Pei, and H. Liu, "Optimization of FOPID controller based on MPFGA," in *Proceedings of the 2019 Chinese Control Conference*, pp. 1846–1850, Guangzhou, China, July 2019.

[4] C. Osinski, G. Villar Leandro, and G. H. Da Costa Oliveira, "Fuzzy PID controller design for LFC in electric power systems," *IEEE Latin America Transactions*, vol. 17, no. 1, pp. 147–154, 2019.

[5] L. Xu, T. Xu, J. Wang, and X. Li, "A fuzzy PID controller-based two-axis compensation device for airborne laser scanning," *IEEE Sensors Journal*, vol. 17, no. 5, pp. 1353–1362, 2016.

[6] R. Sharma, K. P. S. Rana, and V. Kumar, "Performance analysis of fractional order fuzzy PID controllers applied to a robotic manipulator," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4274–4289, 2014.

[7] S. Y. Chen, H. R. Lin, M. C. Yang, and Z. Y. Shen, "Fractional-order fuzzy PID contouring control for a VCMs-based *X-Y* motion stage," in *Proceedings of the 2020 6th International Conference on Control, Automation and Robotics (ICCAR)*, pp. 236–241, Singapore, April 2020.

[8] I. S. Jesus and R. S. Barbosa, "Genetic optimization of fuzzy fractional PD + I controllers," *ISA Transactions*, vol. 57, pp. 220–230, 2015.

[9] T. Mahto and V. Mukherjee, "Fractional order fuzzy PID controller for wind energy-based hybrid power system using quasi-oppositional harmony search algorithm," *IET Generation, Transmission & Distribution*, vol. 11, no. 13, pp. 3299–3309, 2017.

[10] L. Liu, S. Luo, F. Guo, and S. Tan, "Multi-point shortest path planning based on an improved discrete bat algorithm," *Applied Soft Computing*, vol. 95, Article ID 106498, 2020.

[11] X. Chen, B. Xu, C. Mei, Y. Ding, and K. Li, "Teaching-learning-based artificial bee colony for solar photovoltaic parameter estimation," *Applied Energy*, vol. 212, pp. 1578–1588, 2018.

[12] L. F. Grisales-Noreña, O. D. Montoya, and C. Andrés Ramos-Paja, "An energy management system for optimal operation of BSS in DC distributed generation environments based on a parallel PSO algorithm," *Journal of Energy Storage*, vol. 20, Article ID 101488, 2020.

[13] S. Gao, Y. Gao, Y. Zhang, and L. Xu, "Multi-strategy adaptive cuckoo search algorithm," *IEEE Access*, vol. 7, pp. 137642–137655, 2019.

[14] Y. W. Zhang, L. Wang, and Q. D. Wu, "Dynamic adaptation cuckoo search algorithm," *Control and Decision*, vol. 29, no. 4, pp. 617–622, 2014.

[15] D. A. Nugraha, K. Lian, and Suwarno, "A Novel MPPT method based on cuckoo search algorithm and golden section search algorithm for partially shaded PV system," in *Proceedings of the 2018 IEEE Electrical Power and Energy Conference (EPEC)*, pp. 1–6, Toronto, Canada, October 2018.

[16] A. Iglesias, A. Gálvez, P. Suárez et al., "Cuckoo search algorithm with Lévy flights for global-support parametric surface approximation in reverse engineering," *Symmetry*, vol. 10, no. 3, p. 58, 2018.

[17] S. Wu, Z. Jiao, L. Yan, R. Zhang, J. Yu, and C.-Y. Chen, "Development of a direct-drive servo valve with high-frequency voice coil motor and advanced digital controller," *IEEE/ASME Transactions on Mechatronics*, vol. 19, no. 3, pp. 932–942, 2014.

[18] S. Y. Chen and C. Y. Lee, "Digital signal processor based intelligent fractional-order sliding-mode control for a linear voice coil actuator," *IET Control Theory & Applications*, vol. 11, no. 8, pp. 1282–1292, 2017.

[19] D. Huang, V. Venkataramanan, J.-X. Xu, and T. C. T. Huynh, "Contact-induced vibration in dual-stage hard disk drive servo systems and its compensator design," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 8, pp. 4052–4060, 2014.

[20] S. Yabui, T. Atsumi, and T. Inoue, "Coupling controller design for MISO System of head positioning control systems in HDDs," *IEEE Transactions on Magnetics*, vol. 56, no. 5, pp. 1–9, 2020.

[21] C. E. Kim and Y. R. Kim, "Design and analysis of linear voice coil motor for automatic transmission," in *Proceedings of the International Conference on Electrical Machines and Systems (ICEMS)*, College Station, TX, USA, February 2017.

[22] Y.-H. Chang, C.-S. Liu, I.-W. Chen, M.-S. Tsai, and H.-C. Tseng, "Open-loop control of voice coil motor with

magnetic restoring force using high-low frequency composite signals," *IEEE Access*, vol. 7, pp. 146258–146263, 2019.

[23] K. J. Smith, D. J. Graham, and J. A. Neasham, "Design and optimization of a voice coil motor with a rotary actuator for an ultrasound scanner," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 11, pp. 7073–7078, 2015.

[24] R. Wang, X. Yin, Q. Wang, and L. Jiang, "Direct amplitude control for voice coil motor on high frequency reciprocating rig," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 3, pp. 1299–1309, 2020.

[25] S.-Y. Chen, T.-H. Li, and C.-H. Chang, "Intelligent fractional-order backstepping control for an ironless linear synchronous motor with uncertain nonlinear dynamics," *ISA Transactions*, vol. 89, pp. 218–232, 2019.

[26] Akribis Systems Pte Ltd. AVM Series, http://www.akribis-sys.com/.

[27] M. C. Yang, "*Optimal fractional-order PID control for a VCMs-based X-Y motion stage*," M.S. thesis, Department. Electric. Eng., National Taiwan Normal University, Taipei, Taiwan, 2020.

[28] M. Dulău, A. Gligor, and T. M. Dulău, "Fractional order controllers versus integer order controllers," *Procedia Engineering*, vol. 181, pp. 538–545, 2017.

[29] S. Ebrahimkhani, "Robust fractional order sliding mode control of doubly-fed induction generator (DFIG)-based wind turbines," *ISA Transactions*, vol. 63, pp. 343–354, 2016.

[30] G. Zhzo, "Fractional-order fast terminal sliding mode control for a class of dynamical systems," *Mathematical Problems in Engineering*, vol. 2013, Article ID 384921, 10 pages, 2013.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 8863000, 21 pages
https://doi.org/10.1155/2021/8863000

Research Article

Optimization of Transmitter-Receiver Pairing of Spaceborne Cluster Flight Netted Radar for Area Coverage and Target Detection

Tingting Yan,[1,2] **Shengbo Hu**,[1,3] **Jianan Cai**,[1,2] **Jinrong Mo**,[1,3] **and Mingfei Xia**[1,2]

[1]*Institute of Intelligent Information Processing, Guizhou Normal University, Guiyang 550001, China*
[2]*School of Mathematical Sciences, Guizhou Normal University, Guiyang 550001, China*
[3]*Center for RFID and WSN Engineering, Department of Education Guizhou, Guiyang 550001, China*

Correspondence should be addressed to Tingting Yan; 1592624854@qq.com and Shengbo Hu; hsb@nssc.ac.cn

Received 2 September 2020; Revised 17 February 2021; Accepted 23 February 2021; Published 10 March 2021

Academic Editor: Mohamed El Ghami

In this paper, we investigate the optimization problem of the transmitter-receiver pairing of spaceborne cluster flight netted radar (SCFNR) for area coverage and target detection. First of all, we propose the novel concept of SCFNR integrated cluster flight spacecraft with netted radar, the mobility model for bistatic radar pair with twin-satellite mode, and formulate the radar-target distance distribution function and radar-target distance product distribution function with geometric probability method. Secondly, by dividing surveillance region into grids, we define the 0-1 grid coverage matrix for bistatic radar and the transmitter-receiver pairing matrix for SCFNR with using radar equation and the radar-target distance distribution function, and we describe the optimal problem of transmitter-receiver pairing of SCFNR for area coverage and target detection by defining K-grid coverage matrix. Thirdly, we propose a new algorithm integrated particle swarm optimization with Hungarian algorithm (PSO-HA) to address the optimal problem, which is actually one-to-one pairing problem. Finally, we validate the effectiveness and reasonability of the proposed algorithm through numerical analysis.

1. Introduction

As a distributed space system, the spaceborne netted radar is composed of several spatially separated, mutually independent, and cooperative radars in space. Compared with the traditional radar, spaceborne netted radar has advantages of high flexibility, reliability, and antistealth ability [1–3]. In addition, it also has the advantage of being all-weather, wide coverage, and satisfying specific coverage requirements due to its location in outer space [4, 5]. On the other hand, the cluster flight spacecraft has been one of the hot issues regarding the distributed space network, because of its advantages of flexibility, rapid response, low cost, strong scalability, and long lifetime [6–8]. Unlike traditional satellite formation flying applications, cluster flight spacecraft requires nodes to maintain bounded relative distances between tens or hundreds of kilometers and to keep loose geometry for the entire mission lifetime, so that orbit controlling and relative position sensing for the spacecraft can be performed well [6–8]. Some researches are mentioned cluster flight spacecraft. The paper [6] presents cluster-keeping algorithms aimed at minimizing fuel consumption. The paper [9] provides a cooperative control framework aimed at synchronizing the mean-orbital element convergence among cluster-flying satellites. The paper [10] studies the relationship between first docking time and spatial initial distribution and the relationship between first separating time and spatial initial distribution for cluster flight spacecraft. The influence of node transmit power on the QoS performance of cluster flight spacecraft network is analyzed in [11]. In order to improve the performance of cluster flight spacecraft network, the nodal distance distributions are studied in [12]. Hence, we propose the novel concept of SCFNR integrated cluster flight spacecraft with the

spaceborne netted radar, and the optimization problem of SCFNR on coverage is addressed.

Coverage is one of the important issues about radar for target detection, localization, and tracking. According to radar equations, the coverage problem of the spaceborne netted radar is related to many factors such as orbit, antenna gain, transmitted power, and radar cross section. To meet the need of improving the target detection and position, it expects that more radars of SCFNR can cover the surveillance region on the earth, which is completely different from line-of-sight (LOS) coverage of satellite constellation [13, 14]. When the antenna gain, the transmitted power, and radar cross section are constant, the coverage of SCFNR completely depends on the product of transmitter-target and target-receiver distance and the spacecraft orbit. Our previous research shows that the geometry configuration of SCFNR is characterized by high spatiotemporal dynamic and random, which complicates coverage problem of SCFNR. So, the problem about the coverage about SCFNR is more challenging.

The netted radar is a case of multistatic radar [15–19], where transmitters can collaborate with several receivers at different locations. According to the pairing method of transmitter-receiver, the netted radar is mainly classified into three categories: a group of bistatic radars, a single transmitter with several receivers, and a single receiver with several transmitters. To improve the performance of SCFNR, it expects that more radars of SCFNR can cover the surveillance region, and this can be described by the maximum intersection coverage. Actually, the maximum intersection coverage is the classical maximum k-subset intersection (MSI) in graph theory, and it is also a combinatorial optimization problem [20]. To the best of our knowledge, there is not seen much on solving MSI problems. In [21], the authors introduce a GRASP heuristic and propose an integer programming formulation MSI problem. However, to solve the MSI problem about SCFNR is more difficult due to the geometry configuration with spatiotemporal dynamic and random.

To the best of our knowledge, this makes the first paper to investigate the SCFNR coverage problem. The main contributions of our work are summarized as follows:

(1) We propose the novel concept of SCFNR integrated cluster flight spacecraft with netted radar, and the mobility model for bistatic radar pair is established by twin-satellite mode. The distribution function of the product of transmitter-target and receiver-target distance is derived using the method of geometric probability.

(2) According to radar equation, we propose the concept of 0-1 grid coverage matrix for bistatic radar by dividing the surveillance region into grids, and the definition of the transmitter-receiver pairing matrix for SCFNR is given using bistatic radar pairs. These provide an important theoretical basis for optimizing the transmitter-receiver pairing of SCFNR for area coverage and target detection.

(3) We describe the optimal problem of transmitter-receiver pairing of SCFNR for area coverage and target detection by defining K-grid coverage matrix. Also, we propose a new algorithm integrated PSO-HA to address the optimal problem. We validate the effectiveness of the proposed algorithms through numerical calculation.

The rest of the paper is organized as follows: Section 2 reviews the related work. Section 3 proposes the novel concept of SCFNR, establishes mobility model for bistatic radar pair, and derives the distribution function of the product of transmitter-target and receiver-target distance. Section 4 defines the coverage matrix of bistatic radar and pairing matrix of SCFNR based on Section 3 and describes the optimal problem of transmitter-receiver pairing of SCFNR for area coverage and target detection. Section 5 presents the PSO-HA algorithm. Section 6 verifies the effectiveness of the proposed algorithm, and coverage and detection results using numerical calculation are given. Finally, we conclude the paper in Section 7.

2. Related Works

In recent years, with continuing advances in communication technology and micro-electromechanical systems (MEMS) technology, multistatic radar sensing technology has received considerable attention, especially bistatic radar sensing coverage. For instance, in [22], the authors considered the problem of deploying a network of bistatic radars in a region to maximize the worst-case intrusion detectability. They studied the coverage problem of a bistatic radar sensor network and the optimal placement of bistatic radars on a line segment to minimize its vulnerability. In [23], Wang et al. studied the belt barrier coverage with the minimum total placement cost in bistatic radar sensor networks. They proposed a line-based equipartition placement strategy such that all radars placed on a deployment line can form a barrier with some breadth and one or more such placement lines can form a belt barrier with the required breadth. In [24], the authors studied area coverage in bistatic radar sensor networks. They investigated the geometrical relationship between the c-coverage area of a bistatic radar and the distance between its component transmitter and receiver. Then, they reduced the problem dimension by transforming the area coverage problem to point coverage problem by employing the intersection point concept. In [25], the authors studied the worst-case coverage under deterministic deployment, aiming to find optimal deployment locations of radar transmitters and receivers such that the worst-case intrusion detectability was maximized. Then, by developing a novel 2-site Voronoi diagram with graph search techniques, they designed an algorithm to find approximate worst-case intrusion detectability. In [23], the authors studied the belt barrier coverage in bistatic radar sensor networks, which was dependent on the distance between a pair of radar transmitter and receiver. In [26], Wang et al. studied barrier coverage in bistatic radar sensor networks. They formulated the barrier coverage problem as minimum weight barrier coverage problem. By constructing a directed coverage graph, minimum weight barrier coverage problem was transformed into finding k node-disjoint shortest paths. Next, they proposed an energy-efficient algorithm to solve the problem within polynomial.

In addition, intelligent coverage becomes a practical research topic in dynamic sensors network. The paper [27] provided wireless signal coverage schemes for point-to-point and point-to-region and determined the required horizontal rotation angle and pitch rotation angle of the directional antenna intelligent coverage. In [28], the authors established the mobile sensor non-cooperative game model. Then, a local information-based topology control (LITC) algorithm based on this model was proposed, in which sensors move to enhance coverage by exchanging information with neighbors. Also, the application of PSO is wide in complex system. In [29], a systematic data-driven adaptive neuro-fuzzy inference system (ANFIS) modelling methodology was proposed, and a high-performance PSO-LSE method was developed to improve the structure and to identify the consequent parameters of ANFIS model. In [30], the authors proposed an algorithm combined with belief-desire-intention agent with a quantum-behaved particle swami optimization (QPSO) algorithm to optimize a marine generator excitation controller, and the QPSO algorithm was highly robust because its performance was insensitive to the accuracy of system parameters. For intelligent coverage in complex dynamic environment, in [31], a novel trajectory scheduling method based on coverage rate for multiple mobile sinks was presented, especially for large-scale wireless sensor networks, and an improved PSO combined with mutation operator was introduced to search the parking positions with optimal coverage rate. Predictably, considering complex dynamic sensor network, PSO in intelligent coverage is helpful.

For a long time, some works on spaceborne radar coverage are mainly focused on optimizing the orbital design according to the observation and detection requirements. For example, in [32], the authors presented a feasibility analysis of a spaceborne bistatic radar mission for soil moisture retrieval, and they studied the assessment of the spatial coverage from orbital design. In [33], the authors adopted bistatic geometry from space platforms, and they implemented bistatic synthetic aperture radar observation. In [34], based on the analysis of radar cross-section (RCS) characteristic of geostationary orbital targets, the orbital altitude and revisiting period of space-based radar was designed in detail, and they discussed the relationship between image's resolutions of spaceborne inverse synthetic aperture radar and system parameters. In [35], the authors established a spaceborne-airborne bistatic radar model, and then they analyzed moving target detecting performance of the space-time adaptive processing technology.

For cluster flight spacecraft, related researches focus more on orbital control and node connection. The paper [6] presented a methodological development of cluster flight algorithms for disaggregated satellite systems in low Earth orbits. To obtain distance-bounded relative motion, a new constraint on the initial conditions of the modules was developed. In [9], the authors developed the implementable cluster flight-control methods with realistic orbital and actuator modelling. They offered two distributed orbit control laws with fixed-magnitude thrust for satellite cluster flight based on mean-orbital elements. Recently, the team of this paper has done some works on the cluster flight spacecraft network. For example, in [10], the authors proposed the constraint condition of orbital elements for noise-limited fractionated spacecraft network percolating and path formation time. The numerical results showed that the network topology for fractionated spacecraft is time varying and dynamic. The paper [11] investigated the transmit power allocation problem to minimize the average packet error rate at the access point in the cluster flight spacecraft network. Due to the complexity of the calculation, the probability density function of the distance between nodes was fitted using eighth-order polynomial.

3. The Concept of SCFNR

As mentioned above, the spaceborne netted radar is composed of several spatially separated, mutual independent, and cooperative radars in space. The spaceborne netted radar has the advantages of high flexibility, reliability, and anti-stealth ability. In addition, it also has the advantage of being all-weather, wide coverage, and satisfying specific coverage. On the other hand, the cluster flight spacecraft has many advantages such as flexibility, rapid response, low cost, strong scalability, and long lifetime. What's more, cluster flight spacecraft can perform orbit controlling and relative position sensing easily. Hence, we propose the novel concept of SCFNR integrated advantages of both cluster flight spacecraft and the spaceborne netted radar.

Generally, netted radar has the following three cases: (1) a group of bistatic radars, where the output of the bistatic radars are processed centrally to obtain a decision regarding the presence of a target and to estimate parameters. In this case, it is assumed that the transmitters do not interfere with each other, which is typically achieved either by using separate frequency bands or orthogonal transmitted waveforms. At the same time, each receiver is assumed to be able to receive the signals from each transmitter; (2) a single transmitter with several receivers, typically in the case of a high-value unit equipped with the transmitter, for instance an airborne warning and control system, and receivers cooperating to achieve the detection; (3) a single receiver with several transmitters, where a single receiver receives waveforms from several transmitters in different frequency bands to information fusion.

In this paper, we adopt SCFNR with bistatic radar pairs. It is assumed that one-to-one pairing method is taken by SCFNR in any slot of the orbital hyperperiod. So, we assume that each pair of transmitter and receiver can potentially form a bistatic radar. We further assume that orthogonal transmissions are used for interference avoidance. In view of this, we assume that one transmitter can only be connected to one receiver, and the corresponding bistatic radar is formed in any slot of the orbital hyperperiod. Therefore, given a SCFNR consisting of N radars, if N is even, then the pairing of bistatic radars is $N/2$ pairs, and if N is odd, then the pairing of bistatic radars is $(N-1)/2$ pairs and a monostatic radar. Since the monostatic radar can be considered as a bistatic radar with a baseline length 0, it can also be considered that $(N+1)/2$ pairs of bistatic radars is formed.

Based on the above, this paper focuses on the optimization problem of the transmitter-receiver pairing of SCFNR for area of interest coverage and target detection, that is, how to pair transmitter-receiver properly to satisfy the requirements of area coverage and detection in any slot of the orbital hyperperiod. First, the mobility model for SCFNR is

presented and analyzed. Of course, the mobility model for bistatic radar pair can be given, and also the distribution function of the product of transmitter-target and receiver-target distance needs to be derived.

3.1. The Mobility Model for Bistatic Radar Pair. To accomplish the cluster flight model within bounded distance, the twin-satellite model is adopted to study the mobility model for bistatic radar pair. As shown in Figure 1, the transmitter or receiver position is uniformly distributed on sphere within $(M-m)/4$. M is the upper bound of transmitter-receiver distance in SCFNR, and m is the lower bound.

Based on orbit dynamics theory, the orbital hyperperiod can be divided into $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{\mathcal{F}}$, times for fractionated spacecraft [7, 36]. So, there are $\mathcal{T}$ time slots in an orbital period. The orbital hyperperiod is $\mathcal{H} = (\mathcal{T}_{\mathcal{F}} - \mathcal{T}_0)$, time slot $\sigma_k = [\mathcal{T}_{k-1}, \mathcal{T}_k)\,(k = 1, 2, \ldots, \mathcal{T})$ [7]. So, the mobility model of SCFNR can be defined as follows.

Definition 1. In earth-centered inertial (ECI) coordinates, if the position set of N transmitters and receivers in CFSNR is $S(0) = \{S_1(0), S_2(0), \ldots, S_N(0)\}$ at initial time $\mathcal{T}_0$, the position set is $\mathbf{S}(k) = \{S_1(k), S_2(k), \ldots, S_N(k)\}$, and the positions are uniformly distributed within sphere $B(S_i(0), R)\,(i = 1, 2, \ldots, N)$ at time k, where $S_i(0)$ and $R = (M-m)/4$ are the center and radius of the sphere, respectively. Moreover, positions among all transmitters and receivers are mutually independent and independent of all previous locations.

3.2. The Probability Distribution Function of the Distance Product. We consider a SCFNR scenario as shown in Figure 2. Let $\mathbf{T}$ be the transmitter set and $\mathbf{R}$ be the receiver set. Transmitters and receivers are located at different locations. We use $\mathbf{TR}$ to denote all transmitter-receiver pairs. If transmitter $T_i \in \mathbf{T}$ and receiver $R_j \in \mathbf{R}$ choose the same channel, then the bistatic radar $T_iR_j \in \mathbf{TR}$ is formed by T_i and R_j, and different channels can be considered as orthorhombic channels to avoid interference. Without ambiguity, in any time slot of orbital hyperperiod for SCFNR, the position of transmitter and receiver is denoted by S_{iT} and S_{jR}, respectively, where $i \neq j$.

Thus, in SCFNR scenario, let P be a target position in the surveillance region. According to [37], for a bistatic radar $T_iR_j \in \mathbf{TR}$, the signal-to-noise ratio (SNR) of P can be given as

$$\mathrm{SNR} = \frac{K_B}{\left\|S_{iT}P\right\|^2 \left\|PS_{jR}\right\|^2}, \tag{1}$$

where $\|S_{iT}P\|$ and $\|PS_{jR}\|$ denote transmitter-target and target-receiver distances, respectively. K_B is a constant related to the physical-layer parameters of the bistatic radar, such as transmit power, antenna gains of transmitter and receiver, and radar cross-section. However, we are not interested in the abovementioned physical-layer parameters, but transmitter-target and target-receiver distances. For convenience, we assume that the constant is identical for any bistatic radar, i.e., homogeneous bistatic radar also.

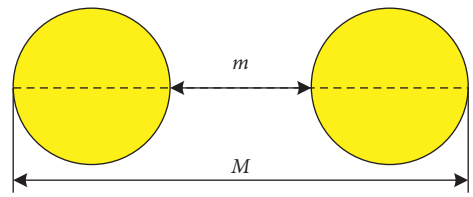


Figure 1: The mobility model for bistatic radar pair.

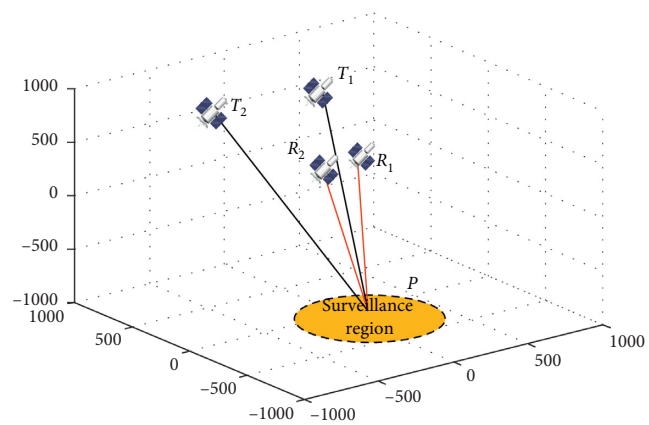


Figure 2: The SCFNR scenario.

As seen from equation (1), the SCFNR performance is determined by $\|S_{iT}P\|\|PS_{jR}\|$, i.e., the product of transmitter-target and target-receiver distances. According to Definition 1, the product is random. Therefore, we need to analyze its distribution.

For convenience, the 2D scenario about transmitter T_i and target P in SCFNR is described in Figure 3. T_i is assumed to be uniformly located in a circle of the two-dimensional plane, and the P is assumed to be fixed. In Figure 3, let $\|S_{iT}P\| = d_i$, h_i be the distance between P and initial orbital position of T_i, where $d_i \in [h_i - R, h_i + R]\,(h_i > R)$. Actually, if P is a target position of earth surface, then h_i can be considered as the orbit height of T_i at initial time.

Therefore, the transmitter-target distance d_i has the distance function given by probability distribution, that is,

$$F_{D_i}(d_i) \triangleq P\{D_i \le d_i\}, \quad h_i - R \le d_i \le h_i + R, h_i > R. \tag{2}$$

Here, $F_{D_i}(d_i)$ is calculated with geometric probability method [38, 39].

Now, we extend the 2D scenario in Figure 3 into 3D scenario. Let Ω be the sphere O and C_0 be the intersection volume between sphere O and the sphere of radius d_i centered at P. Equation (2) can be rewritten as

$$F_{D_i}(d_i) = \frac{\mu(C_0)}{\mu(\Omega)}, \tag{3}$$

where $\mu(\Omega) = 4\pi R^3/3$ is the measure of Ω.

In order to calculate $F_{D_i}(d_i)$, $\mu(C_0)$ can be divided into two cases: (1) $d_i \in [h_i - R, h_i)$; (2) $d_i \in d_i \in [h_i, h_i + R]$. Thus, Theorem 1 about distribution function of d_i can be proved.

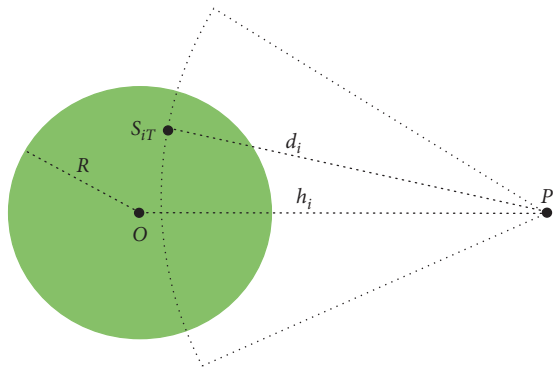


Figure 3: The transmitter-target distance distribution in SCFNR.

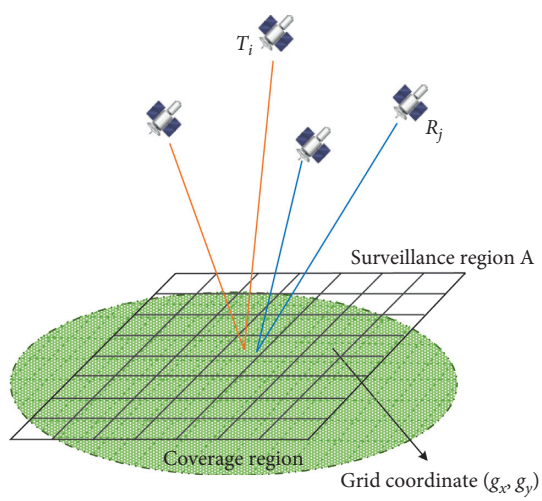


Figure 4: Sketch of coverage region in SCFNR scenario.

Theorem 1. *In SCFNR, if the initial position of transmitter $S_{iT}(0)$ and mobility model $M(t)$ are given, then the distribution function of random variable d_i is*

$$F_{D_i}(d_i) = \begin{cases} 0, & d_i < h_i - R, \\ \dfrac{2d_i^3 + h_a^3 - 3h_a d_i^2 + 2R^3 + h_b^3 - 3h_b R^2}{4R^3}, & h_i - R \le d_i < h_i, \\ \dfrac{2d_i^3 + h_a^3 - 3h_a d_i^2 + 2R^3 - h_c^3 + 3h_c R^2}{4R^3}, & h_i \le d_i \le h_i + R, \\ 1, & h_i + R < d_i, \end{cases} \tag{4}$$

where d_i is the radar-target distance, $h_a = (h_i^2 - R^2 + d_i^2/2h_i)$, $h_b = (h_i^2 + R^2 - d_i^2/2h_i)$, and $h_c = (-h_i^2 - R^2 + d_i^2/2h_i)$.

Proof of Theorem 1. The proof of Theorem 1 is given in Appendix A.

Similarly, if $\|PS_{jR}\| = d_j$ is the distance between target P and receiver R_j, then $\|S_{iT}P\|\|PS_{jR}\| = d_i d_j = d_{ij}$ is the product of transmitter-target and target-receiver distances. Since d_i and d_j are independent, Theorem 2 about distribution function of d_{ij} can be proved. □

Theorem 2. *In SCFNR, if the initial positions of transmitter $S_{iT}(0)$, receiver $S_{jR}(0)$, and mobility model $M(t)$ are given, then distribution function of random variable d_{ij} is*

$$F_{D_{ij}}(d_{ij}) = \begin{cases} 0, & d_{ij} < (h_1 - R)(h_2 - R), \\ F_{D_{ij}}^1(d_{ij}), & (h_1 - R)(h_2 - R) \le d_{ij} < (h_1 + R)(h_2 - R), \\ F_{D_{ij}}^2(d_{ij}), & (h_1 + R)(h_2 - R) \le d_{ij} < (h_1 - R)(h_2 + R), \\ F_{D_{ij}}^3(d_{ij}), & (h_1 - R)(h_2 + R) \le d_{ij} \le (h_1 + R)(h_2 + R), \\ 1, & \text{otherwise}, \end{cases} \tag{5}$$

where $h_1 = \max\{h_i, h_j\}$, $h_2 = \min\{h_i, h_j\}$, and h_j is the orbit height of radar receiver R_j at initial time,

$$
\begin{aligned}
F^1_{D_{ij}}(d_{ij}) &= \frac{C^1_1 d^2_{ij}}{2} + \frac{C^1_2 d^3_{ij}}{3} + \frac{C^1_3 d^4_{ij}}{4} + \frac{C^1_4 d^2_{ij}}{4}\left(2\ln\frac{d_{ij}}{(h_1-R)(h_2-R)} - 1\right) + \frac{C^1_5 d^3_{ij}}{9}\left(3\ln\frac{d_{ij}}{(h_1-R)(h_2-R)} - 1\right)\frac{C^1_6 d^4_{ij}}{16}\left(4\ln\frac{d_{ij}}{(h_1-R)(h_2-R)} - 1\right) + C^1_0,\\
F^2_{D_{ij}}(d_{ij}) &= \frac{C^2_1 d^2_{ij}}{2} + \frac{C^2_2 d^3_{ij}}{3} + \frac{C^2_3 d^4_{ij}}{4} + C^2_0,\\
F^3_{D_{ij}}(d_{ij}) &= \frac{C^3_1 d^2_{ij}}{2} + \frac{C^3_2 d^3_{ij}}{3} + \frac{C^3_3 d^4_{ij}}{4} + \frac{C^3_4 d^2_{ij}}{4}\left(2\ln\frac{d_{ij}}{(h_1+R)(h_2+R)} - 1\right) + \frac{C^3_5 d^3_{ij}}{9}\left(3\ln\frac{d_{ij}}{(h_1+R)(h_2+R)} - 1\right) + \left(3\ln\frac{d_{ij}}{(h_1+R)(h_2+R)}\right) + \frac{C^3_6 d^4_3}{16}\left(4\ln\frac{d_{ij}}{(h_1+R)(h_2+R)} - 1\right) + C^3_0,\\
C^1_1 &= \frac{9(h_1-R)(h_2-R)\left(3h_1h_2 + 2h_1R + 2h_2R + R^2\right)}{16h_1h_2R^6},\\
C^1_2 &= -\frac{9(h_1+h_2)}{4h_1h_2R^5},\\
C^1_3 &= \frac{1}{16h_1h_2R^6}\left(\frac{-27h_1+9R}{2h_1-2R} + \frac{-27h_2+9R}{2h_2-2R}\right),\\
C^1_4 &= \frac{9\left(h_1^2-R^2\right)\left(h_2^2-R^2\right)}{16h_1h_2R^6},\\
C^1_5 &= \frac{9}{4R^6},\\
C^1_6 &= \frac{9}{16h_1h_2R^6},\\
C^1_0 &= -\frac{(h_1-R)^3(h_2-R)^3(h_1+3R)(h_2+3R)}{256h_1h_2R^6},\\
C^2_1 &= \frac{9}{16h_1h_2R^6}\left(\left(h_1^2-R^2\right)\left(h_2^2-R^2\right)\ln\frac{(h_1+R)}{(h_1-R)} - 2Rh_1\left(h_2^2-R^2\right)\right),\\
C^2_2 &= \frac{9}{4R^6}\ln\frac{(h_1+R)}{(h_2-R)} - \frac{18}{4h_1R^5},\\
C^2_3 &= \frac{1}{16h_1h_2R^6}\left(\frac{9R-27h_1}{2h_1-2R} + \frac{9R+27h_2}{2h_2+2R} + 9\ln\frac{(h_1+R)}{(h_1-R)}\right),\\
C^2_0 &= F^1_{D_{ij}}((h_1+R)(h_2-R)) - \frac{C^2_1(h_1+R)^2(h_2-R)^2}{2} - \frac{C^2_2(h_1+R)^3(h_2-R)^3}{3} - \frac{C^2_3(h_1+R)^4(h_2-R)^4}{4},\\
C^3_1 &= -\frac{9(h_1+R)(h_2+R)\left(3h_1h_2 - 2h_1R - 2h_2R + R^2\right)}{16h_1h_2R^6},\\
C^3_2 &= -\frac{9(h_1+h_2)}{4h_1h_2R^5},\\
C^3_3 &= \frac{1}{16h_1h_2R^6}\left(\frac{27h_1+9R}{2h_1+2R} + \frac{27h_2+9R}{2h_2+2R}\right),\\
C^3_4 &= -\frac{9\left(h_1^2-R^2\right)\left(h_2^2-R^2\right)}{16h_1h_2R^6},\\
C^3_5 &= -\frac{9}{4R^6},\\
C^3_6 &= -\frac{9}{16h_1h_2R^6},\\
C^3_0 &= F^2_{D_{ij}}((h_1-R)(h_2+R)) - \frac{C^3_1(h_1-R)^2(h_2+R)^2}{2} - \frac{C^3_2(h_1-R)^3(h_2+R)^3}{3} - \frac{C^3_3(h_1-R)^4(h_2+R)^4}{4}\\
&\quad - \frac{C^3_4 d^2_{ij}}{4}\left(2\ln\frac{(h_1-R)}{(h_1+R)} - 1\right) + \frac{C^3_5 d^3_{ij}}{9}\left(3\ln\frac{(h_1-R)}{(h_1+R)} - 1\right) + \frac{C^3_6 d^3_{ij}}{16}\left(3\ln\frac{(h_1-R)}{(h_1+R)} - 1\right).
\end{aligned}
\tag{6}
$$

Proof of Theorem 2. the proof of Theorem 2 is given in Appendix A. □

4. The Area of Interest Coverage of SCFNR

According to equation (1), the larger the product of transmitter-target and target-receiver distance $\|S_{iT}P\|\|PS_{jR}\|$, the smaller the received SNR, and the probability of the P detected by transmitter-receiver is smaller too. Conversely, the smaller the $\|S_{iT}P\|\|PS_{jR}\|$, the larger the probability will be.

So, we define point coverage of SCFNR for target detection.

Definition 2. Given a threshold value c and a point target P, if there exists a bistatic radar $T_iR_j \in \mathbf{TR}\,(i \neq j)$ and the product of transmitter-target and target-receiver distances in any time slot of orbital hyperperiod is satisfied

$$\|S_{iT}P\|\|PS_{jR}\| < c, \tag{7}$$

then the bistatic radar $T_iR_j \in \mathbf{TR}$ can provide point coverage to point P.

For the sake of analysis, according to the idea of grid, the surveillance region is divided into grids with equal borders, the border length of the grid is able to be elected in accordance with the range resolution of radar. That is, in ECI coordinate, the surveillance region of interest $\mathbf{A}$ (see Figure 4) is encoded in accordance with horizontal encoding $g_x\,(1 \leq g_x \leq N_x)$ and vertical encoding $g_y\,(1 \leq g_y \leq N_y)$; the grid coordinate (g_x, g_y) is denoted by $A_{g_xg_y}$. Thus, the region of interest $\boldsymbol{A}$ can be determined uniquely by all grids and expressed as follows:

$$\mathbf{A} = \left\{A_{g_xg_y} | 1 \leq g_x \leq N_x, 1 \leq g_y \leq N_y\right\}. \tag{8}$$

So, for a bistatic radar $T_iR_j \in \mathbf{TR}$, SNR of each grid can be given as follows:

$$\mathrm{SNR}_{g_xg_y} = \frac{K_B}{\left\|S_{iT}A_{g_xg_y}\right\|^2\left\|A_{g_xg_y}S_{jR}\right\|^2}. \tag{9}$$

Let Γ be the SNR threshold, then $c = \sqrt{K_B/\Gamma}$, and a grid target is covered by a bistatic radar $T_iR_j \in \mathbf{TR}$, if $\mathrm{SNR}_{g_xg_y} \geq \Gamma$. Then, the definition of 0-1 grid coverage matrix on SCFNR can be described as follows:

Definition 3. For a bistatic radar $T_iR_j \in \mathbf{TR}$ in SCFNR, given $A_{g_xg_y} \in \mathbf{A}$, the 0-1 grid coverage matrix is denoted by $\mathbf{U}_{ij} = [u_{ij,g_xg_y}]_{N_x \times N_y}$, where

$$u_{ij,g_xg_y} = \begin{cases} 1, & F_{D_{ij}}\left(\left\|S_{iT}A_{g_xg_y}\right\|\left\|A_{g_xg_y}S_{jR}\right\|\right) \leq F_{D_{ij}}(c), \\ 0, & \text{others}. \end{cases} \tag{10}$$

If $u_{ij,g_xg_y} = 1$ in equation (10), it indicates that the grid $A_{g_xg_y}$ can be covered by the bistatic radar $T_iR_j \in \mathbf{TR}$.

Additionally, to analyze the impact of transmitter-receiver pairs on coverage, 0-1 pairing matrix, which describes the transmitter-receiver pairs selected in SCFNR, can be defined as follows.

Definition 4. For SCFNR, suppose the cardinalities both $\mathbf{T}$ and $\mathbf{R}$ are N. If T_i and R_j are selected as a bistatic radar, let $m_{ij} = 1$; otherwise, $m_{ij} = 0$. Then, 0-1 pairing matrix of transmitter-receivers is denoted by $\mathbf{M} = [m_{ij}]_{N \times N}$.

Note that same grids may be covered by different bistatic radar pairs. Thus, based on Definitions 2 and 4, the introduction of cumulative coverage times $w_{g_xg_y}$ describes the coverage level of SCFNR at grid $A_{g_xg_y}$, that is,

$$w_{g_xg_y} = \sum_{i=1}^{N}\sum_{j=1}^{N} m_{ij}u_{ij,g_xg_y}. \tag{11}$$

As seen from equation (11), $w_{g_xg_y} \in \{0, 1, \ldots, N\}$. On the basis of this, K-grid coverage matrix of SCNFR can be defined as follows.

Definition 5. For SCFNR, given a value $K\,(K \leq N)$, if the variable $c_{g_xg_y}$ is satisfied as

$$c_{g_xg_y} = \begin{cases} 1, & w_{g_xg_y} \geq K, \\ 0, & \text{otherwise}, \end{cases} \tag{12}$$

then the matrix $\mathbf{C} \in \mathbb{R}^{N_x \times N_y}$; $\mathbf{C} = [c_{g_xg_y}]$ is called K-grid coverage matrix.

In equation (12), the total number of elements with 1 in $\mathbf{C}$ represents the grid number satisfying K-grid coverage, and the total number of elements with 0 in $\mathbf{C}$ represents the grid number unsatisfying K-grid coverage, that is,

$$g_0 = N_xN_y - \sum_{g_x=1}^{N_x}\sum_{g_y=1}^{N_y} c_{g_xg_y}. \tag{13}$$

From the point of optimizing system, there is $g_0 \longrightarrow 0$. If the values of g_0 approach 0, then SCFNR can provide completely K-grid coverage to the region $\mathbf{A}$; otherwise, SCFNR fails to provide K-grid coverage to the region $\mathbf{A}$. Therefore, the normalized g_0 is taken as g_1 to measure coverage performance of SCFNR, that is,

$$g_1 = 1 - \frac{1}{N_xN_y}\sum_{g_x=1}^{N_x}\sum_{g_y=1}^{N_y} c_{g_xg_y}. \tag{14}$$

Also, using the radar equation and conditional probability, let $l_{ij} = \|S_{iT}A_{g_xg_y}\|\|A_{g_xg_y}S_{jR}\|$; the detection probability of bistatic radar $T_iR_j \in \mathbf{TR}$ to grid $A_{g_xg_y}$ is given by

$$p_{ij,g_xg_y} = P_r\left\{d_{ij} \geq l_{ij} | d_{ij} \leq c\right\}. \tag{15}$$

Thus, the detection probability of SCFNR radar to grid $A_{g_xg_y}$ is as follows:

$$p_{g_xg_y} = 1 - \prod_{i=1}^{N}\prod_{j=1}^{N}\left(1 - m_{ij}p_{ij,g_xg_y}\right). \tag{16}$$

For the sake of optimization analysis, the worst-case detection probability of all grids is taken as the second objective function to measure the region detection performance of SCFNR, that is,

$$p_{\mathrm{net}} \triangleq \min_{g_x,g_y}\left(p_{g_xg_y}\right). \tag{17}$$

To sum up, in SCFNR coverage scenario, K-grid coverage and detection probability (e.g., g_1 and p_{net}) are functions of pairing matrix $\mathbf{M}$. Therefore, pairing transmitter-receivers with minimum g_1 and maximum p_{net} can be optimized as follows:

$$\begin{aligned}
&\min g_1(\mathbf{M}) = \left(1 - \frac{1}{N_x N_y}\sum_{g_x=1}^{N_x}\sum_{g_y=1}^{N_y} c_{g_x g_y}\right)\\
&\min g_2(\mathbf{M}) = (1 - p_{\text{net}})\\
&\text{s.t.}\quad \begin{aligned} &C_1:\ \textstyle\sum_{i=1}^{N} m_{ij} = 1, \quad \forall j \in \{1,\ldots,N\},\\ &C_2:\ \textstyle\sum_{j=1}^{N} m_{ij} = 1, \quad \forall i \in \{1,\ldots,N\}.\end{aligned}
\end{aligned} \tag{18}$$

The constraints C_1 and C_2 denote that each selected transmitter or receiver can only be associated to one receiver or transmitter. This is actually one-to-one pairing problem [40].

5. Algorithm Design

Obviously, the problem described in equation (18) is a multiobjective optimization problem. Due to the conflicting nature of the two objectives, no solution optimizing all objective functions simultaneously exists in general. Instead, balance among objective functions is taken into account, which is called trade-off analysis in multiobjective optimization, i.e., Pareto optimal solutions [41]. The basic idea is based on a distance measure to determine the solution near by the ideal solution. Here, the weighted Lp norm is taken as the distance measure, that is,

$$g_{\mathbf{M}} = \left[\sum_{i_b=1}^{2} \zeta_{i_b}\left(g_{i_b} - g_{i_b}^*\right)^p\right]^{1/p}, \tag{19}$$

where $g_{i_b}^*$ is the ideal value of i_b-th goal, and $g_1^* = g_2^* = 0$, and ζ_{i_b} is the weight factor of the i_b-th goal.

The problem described in equation (19) is a combinatorial optimization problem. For this problem, particle swarm optimization (PSO) has been proved as an effective tool [42–44]. PSO is based on the behavior of birds flocking [45]. Each particle represents a potential solution to optimization task and all particles fly in the search space to find the optimal solution. But, its solution does not satisfy the constraints C_1 and C_2. As mentioned before, the constraints C_1 and C_2 described in equation (18) are a one-to-one paring problem, which can be solved using Hungarian algorithm (HA) [40]. As a combinatorial optimization method, HA can finish the one-to-one paring task in polynomial time. Therefore, the PSO-HA integrated PSO with HA is proposed. The outline of PSO-HA is given as follows.

Step 1. Initialization.

Suppose swarm size is L, particle is $1\,(1 \le l \le L)$, the maximum number of iterations is $T_{\max}$, iteration time is $t\,(1 \le t \le T_{\max})$, acceleration factors are c_1 and c_2, and the position and velocity of l-th particle are $\overline{M}_l$ and $\mathbf{v}_l$, respectively. Let $t = 1$, and set the parameter values: L, $T_{\max}$, $\overline{M}_l(t)$, $\mathbf{v}_l(t)$, c_1, and c_2.

As a note, $\overline{M}_l$ and $\mathbf{v}_l$ are both $N \times N$ matrixes, in which each element of matrixes is generated randomly. Here, each element of $\overline{M}_l$ and $\mathbf{v}_l$ is set in the range [0, 1] and [−0.5, 0.5], respectively.

Step 2. Update position $\overline{M}_l$ by using PSO.

In each iteration, $\overline{M}_l$ and $\mathbf{v}_l$ are updated as follows:

$$\begin{aligned}
\mathbf{v}_l(t+1) &= \omega(t) \times \mathbf{v}_l(t) + c_1 r_1(t)\left(\boldsymbol{\rho}_l(t) - \overline{M}_l(t)\right) + c_1 r_1(t)\left(\boldsymbol{\rho}_g(t) - \overline{M}_l(t)\right),\\
\overline{M}_l(t+1) &= \overline{M}_l(t) + \mathbf{v}_l(t+1),
\end{aligned} \tag{20}$$

where ρ_l is the current position of l-th particle, ρ_g is the best position of all particles it has visited so far, $\omega(t)$ is the inertia weight which decreases with iteration time as $\omega(t) = 0.9 - 0.5 \times (t/T_{\max})$ [46], and $r_1(t)$ and $r_2(t)$ are random independent variables in the range [0, 1].

Step 3. Generate one-to-one paring matrix $\mathbf{M}_l$ by using HA.

The updating result $\overline{M}_l$ is taken as the cost matrix of l-th particle in HA, and then the optimization problem on one-to-one paring is formulated as follows:

$$\begin{aligned}
&\min \sum_{i=1}^{N}\sum_{j=1}^{N} \overline{m}_{ij}^{l} m_{ij}^{l}\\
&\text{s.t.}\quad \begin{aligned} &C_1:\ \textstyle\sum_{i=1}^{N} m_{ij} = 1, \quad \forall j \in \{1,\ldots,N\},\\ &C_2:\ \textstyle\sum_{j=1}^{N} m_{ij} = 1, \quad \forall i \in \{1,\ldots,N\},\end{aligned}
\end{aligned} \tag{21}$$

where $\overline{m}_{ij}^{l}$ is the element of matrix $\overline{M}_l$, m_{ij}^{l} is the element of matrix $\mathbf{M}_l$, and m_{ij}^{l} is either 0 or 1.

Note that the paring matrix $\mathbf{M}_l$, which satisfies constraints C_1 and C_2, is obtained by using HA.

Step 4. Calculate the fitness function $g_{\mathbf{M}}^{l}$ of l-th particle and determine the best solution $\mathbf{M}_g$ (i.e., optimal paring matrix).

Taking the objective function, i.e., equation (19), as the fitness function of l-th particle (denoted by $g_{\mathbf{M}}^{l}$), the current paring matrix $\mathbf{M}_l$ of l-th particle and optimal paring matrix $\mathbf{M}_g$ of all particles are updated as follows:

$$\mathbf{M}_l(t+1) = \begin{cases} \mathbf{M}_l(t+1), & \text{if } g_{\mathbf{M}}^{l}(t+1) \le g_{\mathbf{M}}^{l}(t), \\ \mathbf{M}_l(t), & \text{others}, \end{cases}$$
$$\mathbf{M}_g(t+1) = \mathbf{M}_l(t+1) \text{ when } \min g_{\mathbf{M}}^{l}(t+1) 1 \le l \le L. \tag{22}$$

Step 5. If $t \le T_{\max}$, then increment t and go to Step 2; otherwise, end.

6. Simulation Analysis

In order to simulate and analyze multiobjective pairing optimization, i.e., coverage and detection performances of SCFNR, in time slot of the orbital hyperperiod, we establish the SCFNR scenario by STK (Satellite Tool Kit) first. Then, we use PSO-HA to find optimal pairing matrix in Windows 10 and MATLAB R2017b environment. At the same time, area of interest coverage and detection probability are analyzed numerically.

6.1. Parameters Setting

6.1.1. Orbital Elements in SCFNR. Suppose SCFNR is composed of 4 pairs of homogeneous bistatic radars. Let $\mathbf{T} = \{T_1, T_2, T_3, T_4\}$, $\mathbf{R} = \{R_1, R_2, R_3, R_4\}$, $m = 30$ km, and $M = 850$ km. According to the orbit design of cluster flight spacecraft proposed in [10], all near circular orbital elements of SCFNR are listed in Table 1.

According to Table 1, all orbital periods can be calculated and are approximated as 6310 seconds using STK, so we believe the orbital hyperperiods of the SCFNR are also 6310 s. In addition, as shown in Figure 5, we can also calculate all relative distances between transmitters and receivers in 172 days by STK. It is observed that the relative distance between any transmitter-receiver always remains below 850 km and above 30 km.

6.1.2. The Target Grid and Other Parameters. Suppose that the longitude and latitude of surveillance regions are in the range [0, 0.07865345] (rad) and [0, 0.07865345] (rad), respectively. That is, surveillance region is set as square with the size of 500 × 500km on the earth surface. The region with longitude and latitude can be divided into $N_x \times N_y$ grids. Let $N_x = 100$ and $N_y = 100$. So, according to coordinate transforming relations between spherical coordinates and rectangular coordinates, each grid can be computed in the ECI coordinate.

For radar equation and PSO-HA, the parameters are listed in Table 2. In this case, the distribution function of d_{ij} for SCFNR in equation (5) can be calculated, as presented in equation (23). At the same time, as shown in Figure 6, we give the curve of distribution function of d_{ij} for SCFNR.

$$F_{D_{ij}}(d_{ij}) = \begin{cases} 0, & d_{ij} < 7.04878 \times 10^5, \\ F_1(d_{ij}), & 7.04878 \times 10^5 \le d_{ij} < 1.04910 \times 10^6, \\ F_2(d_{ij}), & 1.04910 \times 10^6 \le d_{ij} \le 1.56142 \times 10^6 \\ 1, & \text{otherwise}, \end{cases} \tag{23}$$

where

$$\begin{aligned} F_1(d_{ij}) \approx{} & 8.30169 \times 10^{-9} \times d_{ij}^2 - 7.33464 \times 10^{-15} \times \\ & d_{ij}^3 - 6.49148 \times 10^{-21} \times d_{ij}^4 + 3.82234 \times 10^{-9} \times d_{ij}^2 \times \log\left(\frac{d_{ij}}{7.04878 \times 10^5}\right) + 1.01050 \times 10^{-14} \times d_{ij}^3 \times \\ & \log\left(\frac{d_{ij}}{7.04878 \times 10^5}\right) + 1.73646 \times 10^{-21} \times d_{ij}^4 \times \log\left(\frac{d_{ij}}{7.04878 \times 10^5}\right) + 46.52592, \\ F_2(d_{ij}) \approx{} & -1.14225 \times 10^{-8} \times d_{ij}^2 - 5.97942 \times 10^{-16} \times d_{ij}^3 + 5.07373 \times 10^{-21} \times d_{ij}^4 - 3.82234 \times 10^{-9} \times d_{ij}^2 \times \\ & \log\left(\frac{d_{ij}}{1.56143 \times 10^6}\right) - 1.01050 \times 10^{-14} \times d_{ij}^3 \times \log\left(\frac{d_{ij}}{1.56143 \times 10^6}\right) - 1.73646 \times 10^{-21} \times d_{ij}^4 \times \log\left(\frac{d_{ij}}{1.56143 \times 10^6}\right) - 32.88394. \end{aligned} \tag{24}$$

Table 1: All near circular orbital elements of SCFNR.

Parameter	Semimajor axis (km)	Eccentricity (deg)	Inclination (deg)	Argument of perigee (deg)	True anomaly (deg)	Right ascension of ascending node (deg)
T_1	7378.14	0.02	35	0.00000	0.00000	0.00000
T_2	7378.14	0.02	35	0.00163	1.13947	3.38820
T_3	7378.14	0.02	35	0.00068	0.47630	1.82400
T_4	7378.14	0.02	35	359.997	−2.1650	1.70983
R_1	7378.14	0.02	35	0.00106	2.50602	2.50602
R_2	7378.14	0.02	35	0.00022	0.15383	1.08022
R_3	7378.14	0.02	35	359.999	−0.79763	4.02245
R_4	7378.14	0.02	35	0.00344	2.39289	−1.18115

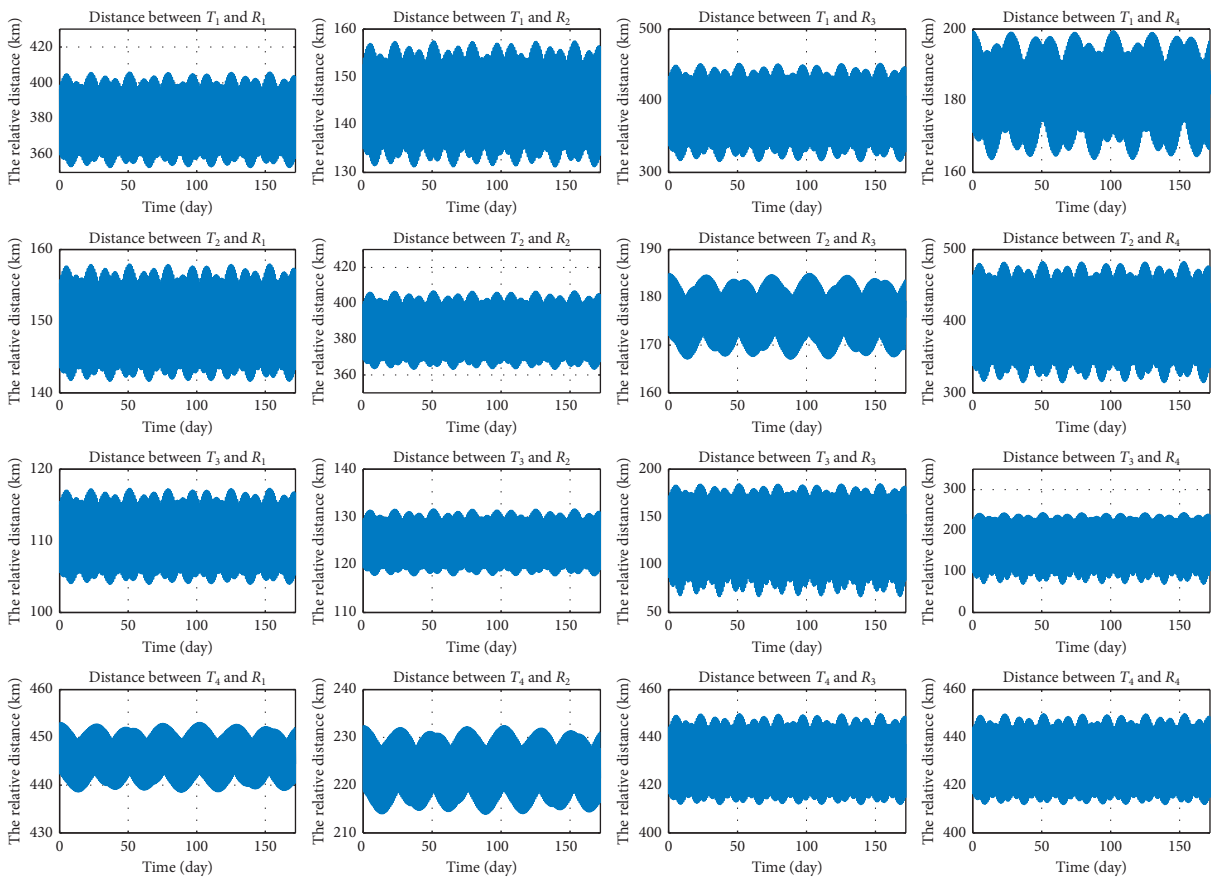


Figure 5: The relative distance between any transmitter-receiver within 172 days.

Table 2: Parameter setting.

Parameter	Value
K_B	$45M^4$
Γ	12.5 dB
h_1	1044.5702 km
h_2	1044.5702 km
c	1.14932×10^6
$F_{D_{ij}}(c)$	0.62269
L	10
$T_{\max}$	100
P	2
c_1	1.49445
c_2	1.49445
R	205 km

Table 3: g_{M} for 24 cases of pairing schemes.

Paring scheme	$\zeta_1 = 0.7$ $\zeta_2 = 0.3$	$\zeta_1 = 0.5$ $\zeta_2 = 0.5$	$\zeta_1 = 0.3$ $\zeta_2 = 0.7$
T_1R_1, T_2R_2, T_3R_3, T_4R_4	0.15540	0.25500	0.35597
T_1R_1, T_2R_2, T_3R_4, T_4R_3	0.16469	0.24026	0.32681
T_1R_1, T_2R_3, T_3R_4, T_4R_2	0.13368	0.19492	0.26509
T_1R_1, T_2R_3, T_3R_2, T_4R_4	0.15069	0.24229	0.33688
T_1R_1, T_2R_4, T_3R_2, T_4R_3	0.15501	0.22166	0.29994
T_1R_1, T_2R_4, T_3R_3, T_4R_2	0.12747	0.19374	0.26614
T_1R_2, T_2R_1, T_3R_3, T_4R_4	0.09853	0.15754	0.21879
T_1R_2, T_2R_1, T_3R_4, T_4R_3	0.11394	**0.13584**	**0.17271**
T_1R_2, T_2R_4, T_3R_3, T_4R_1	0.11183	0.18308	0.25544
T_1R_2, T_2R_4, T_3R_1, T_4R_3	0.11765	0.14436	0.18576
T_1R_2, T_2R_3, T_3R_1, T_4R_4	**0.09844**	0.15016	0.20646
T_1R_2, T_2R_3, T_3R_4, T_4R_1	0.11829	0.18977	0.26374
T_1R_3, T_2R_1, T_3R_2, T_4R_4	0.10855	0.17487	0.24324
T_1R_3, T_2R_1, T_3R_4, T_4R_2	0.10459	0.15033	0.20370
T_1R_3, T_2R_2, T_3R_4, T_4R_1	0.14080	0.23239	0.32475
T_1R_3, T_2R_2, T_3R_1, T_4R_4	0.12640	0.20594	0.28708
T_1R_3, T_2R_4, T_3R_1, T_4R_2	0.13627	0.21010	0.28956
T_1R_3, T_2R_4, T_3R_2, T_4R_1	0.12935	0.21310	0.29770
T_1R_4, T_2R_3, T_3R_2, T_4R_1	0.14810	0.24148	0.33667
T_1R_4, T_2R_3, T_3R_1, T_4R_2	0.11359	0.15552	0.20788
T_1R_4, T_2R_1, T_3R_2, T_4R_3	0.12759	0.16616	0.21857
T_1R_4, T_2R_1, T_3R_3, T_4R_2	0.09847	0.13839	0.18638
T_1R_4, T_2R_2, T_3R_1, T_4R_3	0.13989	0.19168	0.25627
T_1R_4, T_2R_2, T_3R_3, T_4R_1	0.15399	0.25457	0.35585

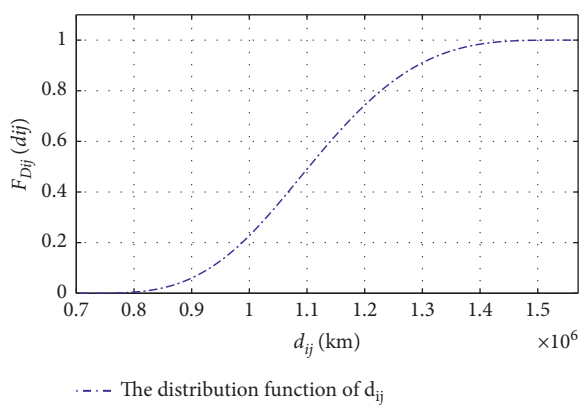


Figure 6: The distance distribution function (equation (23)) associated with d_{ij} in SCFNR.

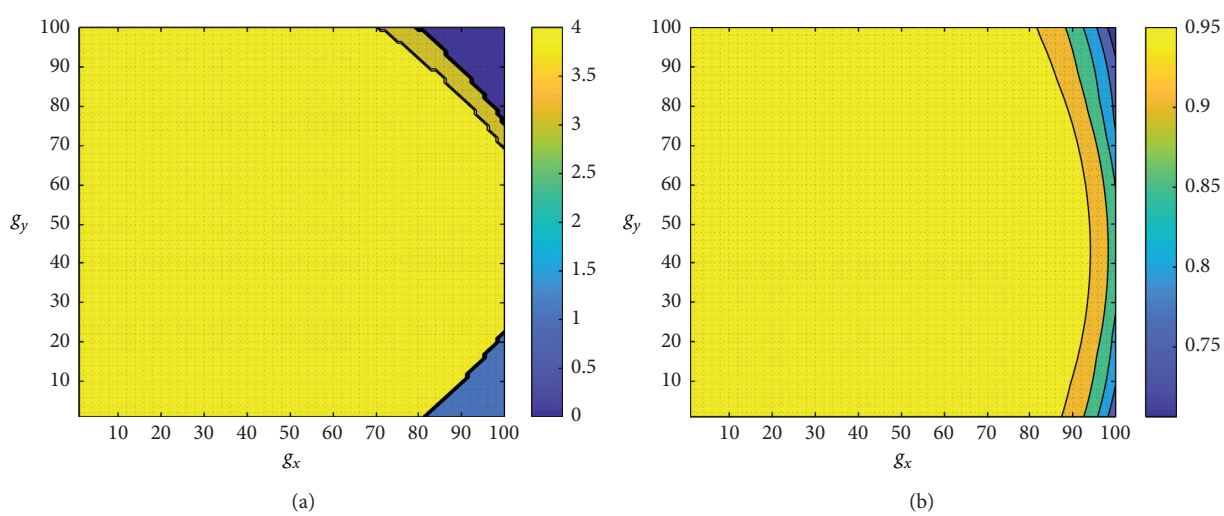


Figure 7: The distributions of coverage level and detection probability with $\zeta_1 = 0.7$, $\zeta_2 = 0.3$, and $\Gamma = 12.5$ dB in grids. (a)$w_{g_x g_y}$. (b)$p_{g_x g_y}$.

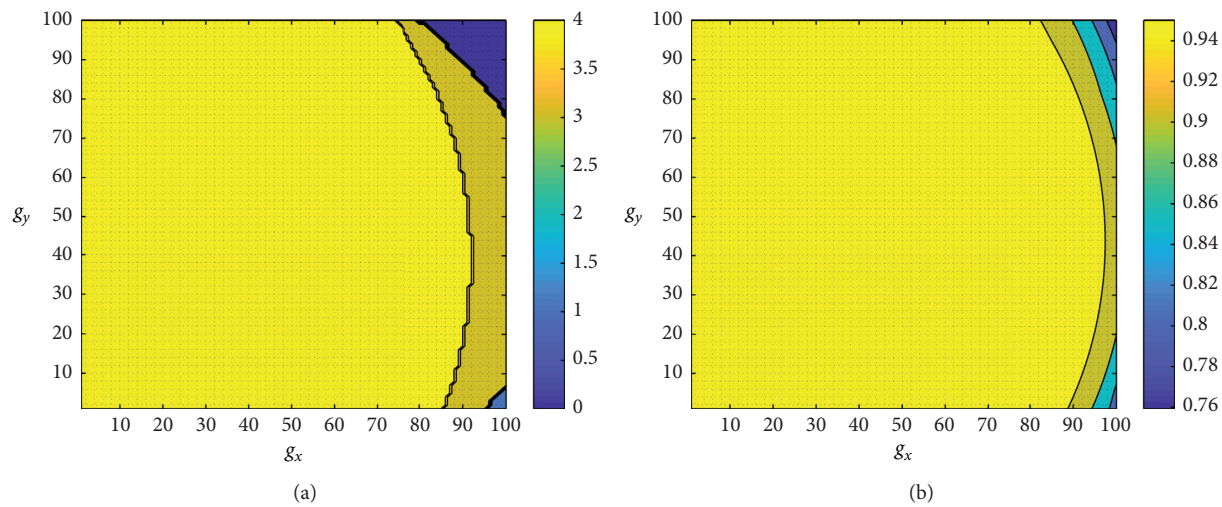


Figure 8: The distributions of coverage level and detection probability with $\zeta_1 = 0.5$, $\zeta_2 = 0.5$, and $\Gamma = 12.5$ dB in grids. (a)$w_{g_x g_y}$. (b)$p_{g_x g_y}$.

6.2. Numerical Result and Analysis

6.2.1. Transmitter-Receiver Pairing Scheme.

Case1: $g_{\mathbf{M}}$ with different weight values in the slot 1 for the 1st orbital hyperperiod.

Considering three conditions, i.e., $\zeta_1 > \zeta_2$, $\zeta_1 = \zeta_2$, and $\zeta_1 < \zeta_2$, we calculate the optimal pairing matrix in the same slot of its orbital hyperperiod under the same simulation environment as described in Section 6.1.

Let $\zeta_1 = 0.7$ and $\zeta_2 = 0.3$; the optimal pairing matrix is given as follows:

$$M_g = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{25}$$

Equation (25) indicates that optimal pairing scheme of bistatic radars is T_1R_2, T_2R_3, T_3R_1, T_4R_4, and $g_{\mathbf{M}} = 0.098349$ with $g_1 = 0.06270$ and $g_2 = 0.29371$.

Let $\zeta_1 = 0.5$ and $\zeta_2 = 0.5$; the optimal pairing matrix is given as follows:

$$\mathbf{M}_g = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{26}$$

Equation (26) indicates that optimal pairing scheme of bistatic radars is T_1R_2, T_2R_1, T_3R_4, T_4R_3, and $g_{\mathbf{M}} = 0.135841$ with $g_1 = 0.1259$ and $g_2 = 0.24075$.

Let $\zeta_1 = 0.3$ and $\zeta_2 = 0.7$; the optimal pairing matrix is given as follows:

$$\mathbf{M}_g = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{27}$$

Equation (27) indicates that the optimal pairing scheme of bistatic radars is T_1R_2, T_2R_1, T_3R_4, T_4R_3, and $g_{\mathbf{M}} = 0.172705$ with $g_1 = 0.1259$ and $g_2 = 0.24075$.

For the sake of comparison, Table 3 lists all pairing schemes for 4 pairs of homogeneous bistatic radars, i.e., 24 cases of pairing schemes and corresponding $g_{\mathbf{M}}$ using the same parameters.

As shown in Table 3, we mark the $g_{\mathbf{M}}$ in bold font. It can be seen that, under conditions of same weight values, the corresponding paring schemes are consistent with equations (25) and (26), respectively. Therefore, PSO-HA is effective and reasonable.

Case 2: $g_{\mathbf{M}}$ with different weight values in the slots 1 and 2 of different orbital hyperperiods.

When $\zeta_1 = 0.7$ and $\zeta_2 = 0.3$, the optimal pairing scheme and corresponding $g_{\mathbf{M}}$ can be listed in Table 4, and when $\zeta_1 = 0.5$ and $\zeta_2 = 0.5$, the optimal pairing scheme and corresponding $g_{\mathbf{M}}$ can be listed in Table 5.

As Tables 4 and 5 show, the optimal pairing schemes are various in different slots of different orbital hyperperiods, and$\zeta_2 = 0.3$ is different as well. From this result, we conclude that geometric topology of SCFNR with high dynamic and random leads to optimal pairing scheme with dynamic and random.

6.2.2. Coverage Level and Detection Probability.

Case 1: when $\Gamma = 12.5$ dB,

Using the optimal pairing matrices given by equations (25)–(27), we calculate the corresponding distributions of coverage level and detection probability as shown in Figures 7 and 8. The coverage level and detection probability corresponding to equations (26) and (27) are the same due to the same optimal pairing matrices in equations (26) and (27). So, we only need to give the two distributions with $\zeta_1 = 0.7$, $\zeta_2 = 0.3$ and $\zeta_1 = 0.5$, $\zeta_2 = 0.5$, respectively.

Table 4: Optimal pairing scheme and corresponding $g_{\mathbf{M}}$ with $\zeta_1 = 0.7$ and $\zeta_2 = 0.3$.

Orbital hyperperiod	Time slot	Optional pairing scheme	$g_{\mathbf{M}}$
The 2nd orbital hyperperiod	1	T_1R_2, T_2R_1, T_3R_3, T_4R_4	0.063586
	2	T_1R_1, T_2R_2, T_3R_3, T_4R_4	0.139745
The 3th orbital hyperperiod	1	T_1R_2, T_2R_1, T_3R_3, T_4R_4	0.047843
	2	T_1R_3, T_2R_4, T_3R_1, T_4R_3	0.123073
The 12th orbital hyperperiod	1	T_1R_2, T_2R_3, T_3R_4, T_4R_1	0.009715
	2	T_1R_1, T_2R_4, T_3R_3, T_4R_2	0.198599
The 16th orbital hyperperiod	1	T_1R_2, T_2R_3, T_3R_4, T_4R_1	0.068867
	2	T_1R_1, T_2R_3, T_3R_4, T_4R_2	0.295543

Table 5: Optimal pairing scheme and corresponding $g_{\mathbf{M}}$ with $\zeta_1 = 0.5$ and $\zeta_2 = 0.5$.

Orbital hyperperiod	Time slot	Optional pairing scheme	$g_{\mathbf{M}}$
The 2nd orbital hyperperiod	1	T_1R_3, T_2R_1, T_3R_4, T_4R_2	0.098536
	2	T_1R_2, T_2R_1, T_3R_3, T_4R_4	0.072687
The 3th orbital hyperperiod	1	T_1R_3, T_2R_4, T_3R_2, T_4R_1	0.156137
	2	T_1R_2, T_2R_4, T_3R_1, T_4R_3	0.156137
The 12th orbital hyperperiod	1	T_1R_2, T_2R_3, T_3R_4, T_4R_1	0.015381
	2	T_1R_1, T_2R_4, T_3R_3, T_4R_2	0.251995
The 16th orbital hyperperiod	1	T_1R_4, T_2R_3, T_3R_1, T_4R_2	0.090544
	2	T_1R_1, T_2R_3, T_3R_4, T_4R_2	0.380717

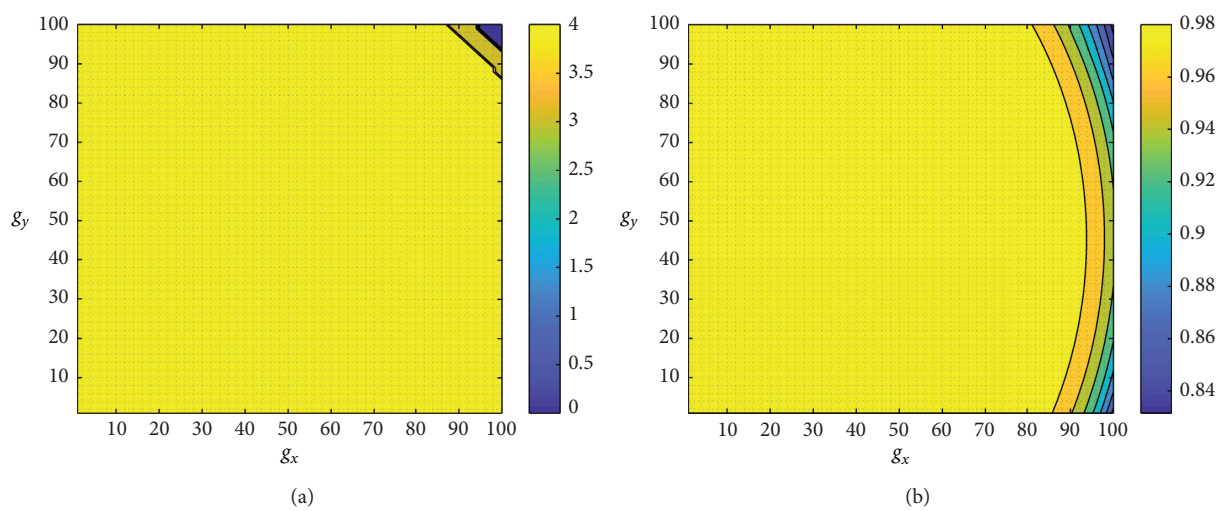


Figure 9: The distributions of coverage level and detection probability with $\zeta_1 = 0.7$, $\zeta_2 = 0.3$, and $\Gamma = 12.0$ dB in grids. (a)$w_{g_x g_y}$. (b)$p_{g_x g_y}$.

In Figures 7 and 8, it is observed that either the distribution of coverage level or the distribution of detection probability is roughly the same. Besides, the higher the coverage level, the higher the detection probability.

Case 2: when $\Gamma = 12.0$ dB,

In this case, keeping other parameters unchanged, the distributions of coverage level and detection probability are shown in Figures 9–11.

In Figures 9–11, it is observed that the three distributions of coverage level and detection probability are roughly the same with different weight values. However, there is considerable difference between the distributions with $\Gamma = 12.5$ dB and $\Gamma = 12.0$ dB, and we find that the SNR threshold has a great influence on coverage level and detection probability in SCFNR. We also observe that the smaller the threshold Γ, the larger the coverage level and detection probability. These show that the proposed PSO-HA, coverage, and detection probability model are reasonable and effective, especially for coverage and detection performance measured by distance function.

In addition, using PSO-HA to solve optimal pairing matrix, considering three weight values, we give the relationship between iteration and $g_{\mathbf{M}}$ in slot 1 for the 1st orbital hyperperiod. As shown in Figure 12, the iterative process has good convergence.

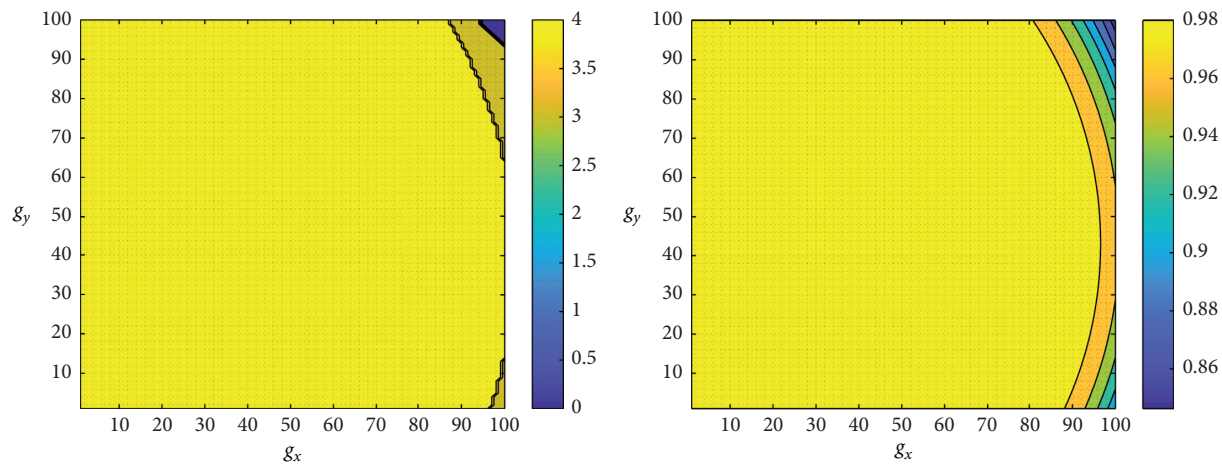


Figure 10: The distributions of coverage level and detection probability with $\zeta_1 = 0.5$, $\zeta_2 = 0.5$, and $\Gamma = 12.0$ dB in grids. (a)$w_{g_x g_y}$. (b)$p_{g_x g_y}$.

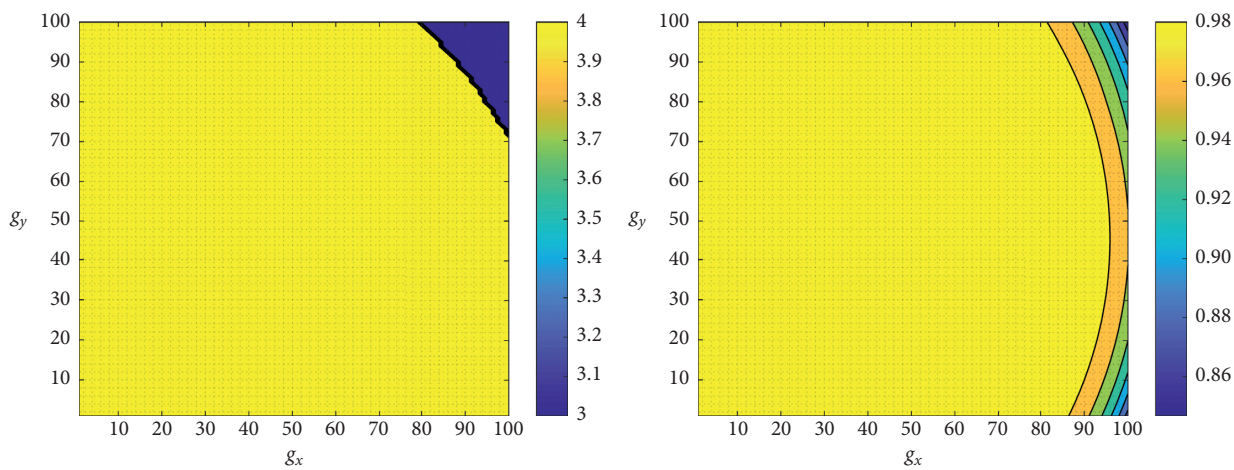


Figure 11: The distributions of coverage level and detection probability with $\zeta_1 = 0.3$, $\zeta_2 = 0.7$, and $\Gamma = 12.0$ dB in grids. (a)$w_{g_x g_y}$. (b)$p_{g_x g_y}$.

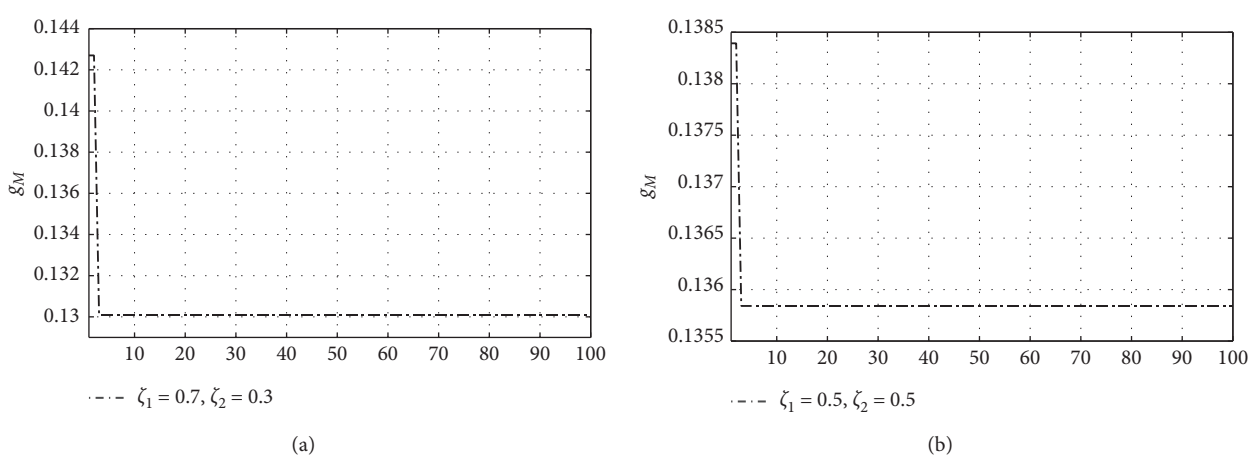
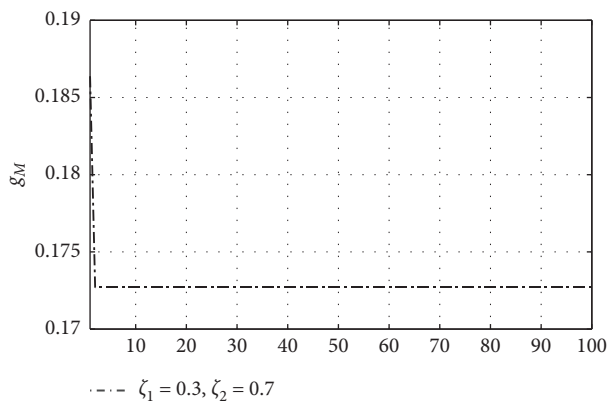
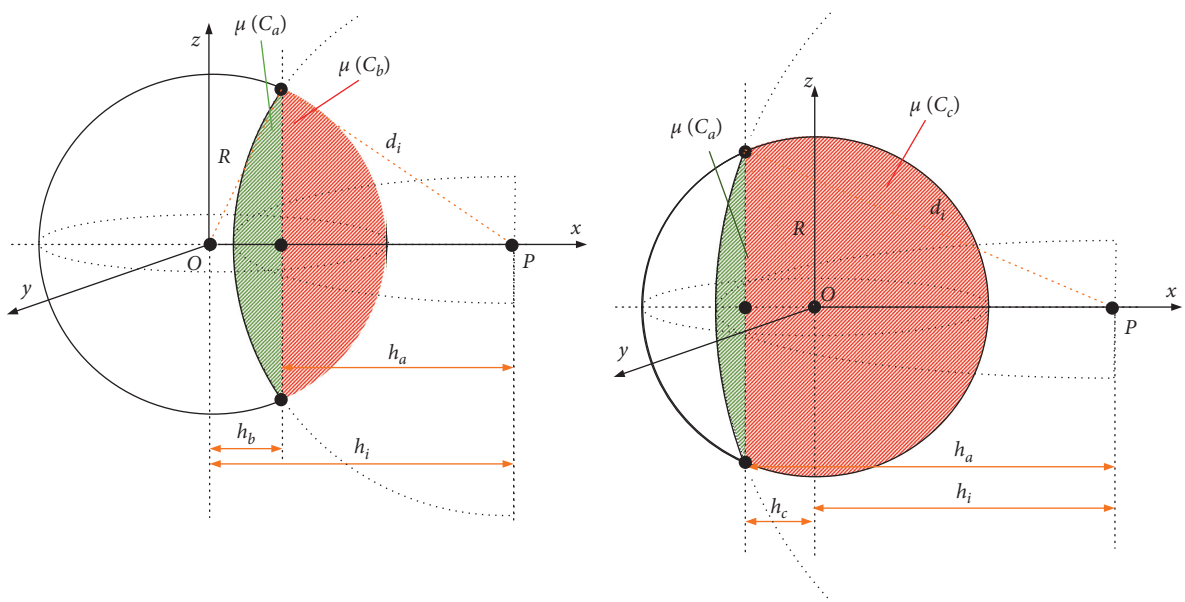


Figure 12: Continued.



(c)

Figure 12: The relationship between iteration and objective function with different weight values.



(a) (b)

Figure 13: The transmitter-target distance distribution in SCFNR. (a)$d_i \in [h_i - R, h_i)$. (b)$d_i \in [h_i, h_i + R]$.

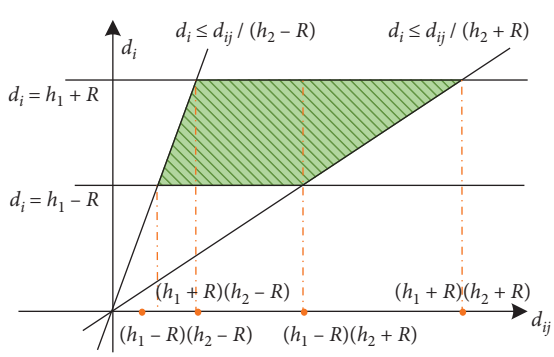


Figure 14: The feasible region with respect to d_i and d_{ij}.

7. Conclusions

In this paper, we study the optimization problem of the transmitter-receiver pairing of SCFNR for area coverage and target detection. Firstly, the novel concept of SCFNR integrated cluster flight spacecraft with netted radar is proposed. By establishing the mobility model for bistatic radar pair with twin-satellite mode, we have derived the radar-target distance distribution function and radar-target distance product distribution function with geometric probability method. Secondly, the radar-target distance distribution function and radar-target distance product distribution function with geometric probability method are proposed; we present the 0-1 grid coverage matrix for the bistatic radar and the transmitter-receiver paring matrix for SCFNR using the radar equation and the radar-target distance distribution function. Next, we describe the optimal problem of transmitter-receiver pairing of SCFNR for area coverage and target detection by defining K-grid coverage matrix. Finally, we propose new PSO-HA for the problem. We validate the effectiveness and reasonability of the proposed algorithm through numerical analysis. The numerical results can also be concluded as follows:

(1) Geometric topology of SCFNR with characteristic great dynamic and random leads to optimal paring scheme with characteristic time varying and random;

(2) When the paring matrix is optimal, the coverage level and detection probability with different weigh values had a slight change;

(3) SNR threshold had a great discernible impact on coverage and detection. In other words, the smaller the SNR threshold, the more achievable its coverage and detection will be.

In the future, to further develop the theory and application of SCFNR, we will study problems on target detection, localization, and tracking.

Appendix

Proof of Theorem 1. For distance distribution $F_{D_i}(d_i)$, we can obtain using equation (3), where $\mu(\Omega) = 4\pi R^3/3$ and intersection volume $\mu(C_0)$ is unknown. In order to calculate $\mu(C_0)$, we categorize as follows:

(1) When $d_i \in [h_i - R, h_i)$, as shown in Figure 13(a), there exists $h_a + h_b = h_i$, $d_i^2 - h_a^2 = R^2 - h_b^2$, thus $h_a = (h_i^2 - R^2 + d_i^2/2h_i)$, $h_b = (h_i^2 + R^2 - d_i^2/2h_i)$, and the intersection volume $\mu(C_0) = \mu(C_a) + \mu(C_b)$. Actually, in Figure 13(a), $\mu(C_a)$ and $\mu(C_b)$ are one part of sphere P with d_i radius and sphere O with R radius, respectively. The methods to calculate $\mu(C_a)$ and $\mu(C_b)$ are the same.

For $\mu(C_a)$, let $x^2 + y^2 + z^2 = d_i^2$ (see Figure 13(a) again); using triple integral, we can calculate it as follows:

$$\mu(C_a) = \int_{h_a}^{d_i} \mathrm{d}x \iint \mathrm{d}y\mathrm{d}z = \int_{h_a}^{d_i} \pi\left(d_i^2 - x^2\right)\mathrm{d}x = \frac{2\pi d_i^3}{3} + \frac{\pi h_a^3}{3} - \pi d_1^2 h_a. \tag{A.1}$$

In the same way, analogous, we can get $\mu(C_b)$, that is,

$$\mu(C_b) = \frac{2\pi R^3}{3} + \frac{\pi h_b^3}{3} - \pi R^2 h_b. \tag{A.2}$$

Hence, we have

$$\mu(C_0) = \frac{2\pi d_i^3}{3} + \frac{\pi h_a^3}{3} - \pi d_i^2 h_a + \frac{2\pi R^3}{3} + \frac{\pi h_b^3}{3} - \pi R^2 h_b. \tag{A.3}$$

Then, substituting equation (A.3) and $\mu(\Omega) = 4\pi R^3/3$ into equation (3), $F_{D_i}(d_i)$ can be calculated as

$$F_{D_i}(d_i) = \frac{2d_i^3 + h_a^3 - 3h_a d_i^2 + 2R^3 + h_b^3 - 3h_b R^2}{4R^3}. \tag{A.4}$$

(2) When $d_i \in [h_i, h_i + R]$, as shown in Figure 13(b), there exists $h_a - h_c = h_i$, $d_i^2 - h_a^2 = R^2 - h_c^2$, thus $h_a = (h_i^2 - R^2 + d_i^2/2h_i)$, $h_c = (-h_i^2 - R^2 + d_i^2/2h_i)$. The intersection volume $\mu(C_0) = \mu(C_a) + \mu(C_c)$. In the same way described in equation (A.1), we can calculate $\mu(C_c)$ as follows:

$$\mu(C_c) = \frac{2\pi R^3}{3} - \frac{\pi h_c^3}{3} + \pi R^2 h_c. \tag{A.5}$$

Hence, we have

$$\mu(C_0) = \frac{2\pi d_i^3}{3} + \frac{\pi h_a^3}{3} - \pi d_i^2 h_a + \frac{2\pi R^3}{3} - \frac{\pi h_c^3}{3} + \pi R^2 h_c. \tag{A.6}$$

Then, substituting equation (A.6) and $\mu(\Omega) = 4\pi R^3/3$ into equation (3), $F_{D_i}(d_i)$ can be calculated as

$$F_{D_i}(d_i) = \frac{2d_i^3 + h_a^3 - 3h_a d_i^2 + 2R^3 - h_c^3 + 3h_c R^2}{4R^3}. \tag{A.7}$$

To sum up, $F_{D_i}(d_i)$ is given by

$$F_{D_i}(d_i) = \begin{cases} 0, & d_i < h_i - R, \\ \dfrac{2d_i^3 + h_a^3 - 3h_a d_i^2 + 2R^3 + h_b^3 - 3h_b R^2}{4R^3}, & h_i - R \le d_i < h_i, \\ \dfrac{2d_i^3 + h_a^3 - 3h_a d_i^2 + 2R^3 - h_c^3 + 3h_c R^2}{4R^3}, & h_i \le d_i \le h_i + R, \\ 1, & h_i + R < d_i. \end{cases} \tag{A.8}$$

Proof of Theorem 2. Let $\|PS_{jR}\| = d_j$ be the distance between target P and receiver, then $\|S_{iT}P\|\|PS_{jR}\| = d_i d_j = d_{ij}$ is the product of transmitter-target and target-receiver distances. To calculate the distribution function of d_{ij}, the accurate probability density functions of d_i and $d_j$are indispensable.

In Theorem 1, we have got the distance function $F_{D_i}(d_i)$, so the probability density function of d_i is

$$f_{D_i}(d_i) = \frac{\partial F_{D_i}(d_i)}{\partial d_i}. \tag{A.9}$$

In the same way, we can also get $f_{D_j}(d_j)$. Since the two random variables d_i and d_j are independent, the probability density function of d_{ij} can be obtained by

$$f_{D_{ij}}(d_{ij}) = \int_{-\infty}^{\infty} \frac{1}{|d_i|} f_{D_i}(d_i) f_{D_j}\left(\frac{d_{ij}}{d_i}\right) \mathrm{d}d_i. \tag{A.10}$$

□

For convenience, let $h_1 = \max\{h_i, h_j\}$ and $h_2 = \min\{h_i, h_j\}$. Obviously, $d_i > 0$; thus, $f_{D_{ij}}(d_{ij})$ can be rewritten as follows:

$$f_{D_{ij}}(d_{ij}) = \int_{-\infty}^{+\infty} \frac{1}{d_i} f_{D_i}(d_i) f_{D_j}\left(\frac{d_{ij}}{d_i}\right) \mathrm{d}d_i, \tag{A.11}$$

where

$$f_{D_i}(d_i) = \begin{cases} \dfrac{3(R^2 - h_1^2)d_i + 6h_1 d_i^2 - 3d_i^3}{4R^3 h_1}, & h_1 - R \le d_i \le h_1 + R, \\ 0, & \text{otherwise}, \end{cases}$$
$$f_{D_j}(d_j) = \begin{cases} \dfrac{3(R^2 - h_2^2)d_j + 6h_2 d_j^2 - 3d_j^3}{4R^3 h_2}, & h_2 - R \le d_j \le h_2 + R, \\ 0, & \text{otherwise}. \end{cases} \tag{A.12}$$

Thus, the distribution function of d_{ij} is

$$\begin{aligned} F_{D_{ij}}(d_{ij}) &= \int_{(h_1-R)(h_2-R)}^{d_{ij}} f_{D_{ij}}(d_{ij}) \mathrm{d}d_{ij} \\ &= \int_{(h_1-R)(h_2-R)}^{d_{ij}} \int_{-\infty}^{+\infty} \frac{1}{d_i} f_{D_i}(d_i) f_{D_j}\left(\frac{d_{ij}}{d_i}\right) \mathrm{d}d_i \mathrm{d}d_{ij}. \end{aligned} \tag{A.13}$$

When equation (A.13) is not equal to 0, the feasible region with respect to d_i and d_{ij} is shown in Figure 14. Therefore, for equation (A.13), we categorize as follows:

(1) When $(h_1 - R)(h_2 - R) \le d_{ij} < (h_1 + R)(h_2 - R)$, we get

$$
\begin{aligned}
F_{D_{ij}}^{1}\left(d_{ij}\right) &= \int_{(h_1-R)(h_2-R)}^{d_{ij}} \int_{(h_1-R)}^{d_{ij}/(h_2-R)} \frac{1}{d_i} f_{D_i}(d_i) f_{D_j}\left(\frac{d_{ij}}{d_i}\right) \mathrm{d}d_i \mathrm{d}d_{ij} \\
&= \int_{(h_1-R)(h_2-R)}^{d_{ij}} \left(C_1^1 d_{ij} + C_2^1 d_{ij}^2 + C_3^1 d_{ij}^3\right) \mathrm{d}d_{ij} + \int_{(h_1-R)(h_2-R)}^{d_{ij}} \left(C_4^1 d_{ij} \ln \frac{d_{ij}}{(h_1-R)(h_2-R)} + C_5^1 d_{ij}^2 \ln \frac{d_{ij}}{(h_1-R)(h_2-R)}\right) \mathrm{d}d_{ij} \\
&\quad + \int_{(h_1-R)(h_2-R)}^{d_{ij}} C_6^1 d_{ij}^3 \ln \frac{d_{ij}}{(h_1-R)(h_2-R)} \mathrm{d}d_{ij} \\
&= \frac{C_1^1 d_{ij}^2}{2} + \frac{C_2^1 d_{ij}^3}{3} + \frac{C_3^1 d_{ij}^4}{4} + \frac{C_4^1 d_{ij}^2}{4}\left(2 \ln \frac{d_{ij}}{(h_1-R)(h_2-R)} - 1\right) \\
&\quad + \frac{C_5^1 d_{ij}^3}{9}\left(3 \ln \frac{d_{ij}}{(h_1-R)(h_2-R)} - 1\right) \\
&\quad + \frac{C_6^1 d_{ij}^4}{16}\left(4 \ln \frac{d_{ij}}{(h_1-R)(h_2-R)} - 1\right) + C_0^1,
\end{aligned}
\tag{A.14}
$$

where

$$
\begin{aligned}
C_1^1 &= \frac{9(h_1-R)(h_2-R)\left(3h_1h_2 + 2h_1R + 2h_2R + R^2\right)}{16h_1h_2R^6}, \\
C_2^1 &= -\frac{9(h_1+h_2)}{4h_1h_2R^5}, \\
C_3^1 &= \frac{1}{16h_1h_2R^6}\left(\frac{-27h_1 + 9R}{2h_1 - 2R} + \frac{-27h_2 + 9R}{2h_2 - 2R}\right), \\
C_4^1 &= \frac{9\left(h_1^2 - R^2\right)\left(h_2^2 - R^2\right)}{16h_1h_2R^6}, \\
C_5^1 &= \frac{9}{4R^6}, \\
C_6^1 &= \frac{9}{16h_1h_2R^6}, \\
C_0^1 &= -\frac{(h_1-R)^3(h_2-R)^3(h_1+3R)(h_2+3R)}{256h_1h_2R^6}.
\end{aligned}
\tag{A.15}
$$

(2) When $(h_1+R)(h_2-R) \le d_{ij} < (h_1-R)(h_2+R)$, we get

$$
\begin{aligned}
F_{D_{ij}}^{2}\left(d_{ij}\right) &= \int_{(h_1+R)(h_2-R)}^{d_{ij}} \int_{(h_1-R)}^{(h_1+R)} \frac{1}{d_i} f_{D_i}(d_i) f_{D_j}\left(\frac{d_{ij}}{d_i}\right) \mathrm{d}d_i \mathrm{d}d_{ij} \\
&= \int_{(h_1+R)(h_2-R)}^{d_{ij}} \left(C_1^2 d_{ij} + C_2^2 d_{ij}^2 + C_3^2 d_{ij}^3\right) \mathrm{d}d_{ij} \\
&= \frac{C_1^2 d_{ij}^2}{2} + \frac{C_2^2 d_{ij}^3}{3} + \frac{C_3^2 d_{ij}^4}{4} + C_0^2,
\end{aligned}
\tag{A.16}
$$

where

$$
\begin{aligned}
C_1^2 &= \frac{9}{16h_1h_2R^6}\left(\left(h_i^2 - R^2\right)\left(h_2^2 - R^2\right)\ln\frac{(h_1+R)}{(h_1-R)} - 2Rh_1\left(h_2^2 - R^2\right)\right), \\
C_2^2 &= \frac{9}{4R^6}\ln\frac{(h_1+R)}{(h_2-R)} - \frac{18}{4h_1R^5}, \\
C_3^2 &= \frac{1}{16h_1h_2R^6}\left(\frac{9R - 27h_1}{2h_1 - 2R} + \frac{9R + 27h_2}{2h_2 + 2R} + 9\ln\frac{(h_1+R)}{(h_1-R)}\right), \\
C_0^2 &= F_{D_{ij}}^{1}\left((h_1+R)(h_2-R)\right) - \frac{C_1^2(h_1+R)^2(h_2-R)^2}{2} - \frac{C_2^2(h_1+R)^3(h_2-R)^3}{3} - \frac{C_3^2(h_1+R)^4(h_2-R)^4}{4}.
\end{aligned}
\tag{A.17}
$$

(3) When $(h_1 - R)(h_2 + R) \le d_{ij} < (h_1 + R)(h_2 + R)$, we get

$$\begin{aligned}
F_{D_{ij}}^3(d_{ij}) &= \int_{(h_1-R)(h_2+R)}^{d_{ij}} \int_{d_{ij}/(h_2+R)}^{(h_1-R)} \frac{1}{d_i} f_{D_i}(d_i) f_{D_j}\left(\frac{d_{ij}}{d_i}\right) \mathrm{d}d_i \mathrm{d}d_{ij} \\
&= \int_{(h_1-R)(h_2+R)}^{d_{ij}} \left(C_1^3 d_{ij} + C_2^3 d_{ij}^2 + C_3^3 d_{ij}^3\right) \mathrm{d}d_{ij} + \\
&\int_{(h_1-R)(h_2+R)}^{d_{ij}} \left(C_4^3 d_{ij} \ln \frac{d_{ij}}{(h_1+R)(h_2+R)} + C_5^3 d_{ij}^2 \ln \frac{d_{ij}}{(h_1+R)(h_2+R)}\right) \mathrm{d}d_{ij} \\
&+ \int_{(h_1-R)(h_2+R)}^{d_{ij}} C_6^3 d_{ij}^3 \ln \frac{d_{ij}}{(h_1+R)(h_2+R)} \mathrm{d}d_{ij} \\
&= \frac{C_1^3 d_{ij}^2}{2} + \frac{C_2^3 d_{ij}^3}{3} + \frac{C_3^3 d_{ij}^4}{4} + \frac{C_4^3 d_{ij}^2}{4}\left(2 \ln \frac{d_{ij}}{(h_1+R)(h_2+R)} - 1\right) + \frac{C_5^3 d_{ij}^3}{9}\left(3 \ln \frac{d_{ij}}{(h_1+R)(h_2+R)} - 1\right) \\
&+ \frac{C_6^3 d_3^4}{16}\left(4 \ln \frac{d_{ij}}{(h_1+R)(h_2+R)} - 1\right) + C_0^3,
\end{aligned} \tag{A.18}$$

where

$$\begin{aligned}
C_1^3 &= -\frac{9(h_1+R)(h_2+R)\left(3h_1h_2 - 2h_1R - 2h_2R + R^2\right)}{16h_1h_2R^6}, \\
C_2^3 &= -\frac{9(h_1+h_2)}{4h_1h_2R^5}, \\
C_3^3 &= \frac{1}{16h_1h_2R^6}\left(\frac{27h_1 + 9R}{2h_1 + 2R} + \frac{27h_2 + 9R}{2h_2 + 2R}\right), \\
C_4^3 &= -\frac{9\left(h_1^2 - R^2\right)\left(h_2^2 - R^2\right)}{16h_1h_2R^6}, \\
C_5^3 &= -\frac{9}{4R^6}, \\
C_6^3 &= -\frac{9}{16h_1h_2R^6}, \\
C_0^3 &= F_{D_{ij}}^2\left((h_1-R)(h_2+R)\right) - \frac{C_1^3(h_1-R)^2(h_2+R)^2}{2} - \frac{C_2^3(h_1-R)^3(h_2+R)^3}{3} \\
&- \frac{C_3^3(h_1-R)^4(h_2+R)^4}{4} - \frac{C_4^3 d_{ij}^2}{4}\left(2 \ln \frac{(h_1-R)}{(h_1+R)} - 1\right) + \frac{C_5^3 d_{ij}^3}{9}\left(3 \ln \frac{(h_1-R)}{(h_1+R)} - 1\right) + \frac{C_6^3 d_{ij}^3}{16}\left(3 \ln \frac{(h_1-R)}{(h_1+R)} - 1\right).
\end{aligned} \tag{A.19}$$

□

Data Availability

The data used to support the findings of this study are described in Section 6.1 of this article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research work was supported by the Guizhou Province Education Department Projects of China (KY[2017]031 and KY[2020]007).

References

[1] H.-Y. Zhao, Z.-J. Zhang, J. Liu, S. Zhou, J. Zheng, and W. Liu, "Target detection based on *F*-test in passive multistatic radar," *Digital Signal Processing*, vol. 79, pp. 1–8, 2018.

[2] Y. Zhao, Y. Zhao, and C. Zhao, "A novel algebraic solution for moving target localization in multi-transmitter multi-receiver passive radar," *Signal Processing*, vol. 143, pp. 303–310, 2018.

[3] X. Zhang, H. Li, and B. Himed, "Multistatic detection for passive radar with direct-path interference," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 53, no. 2, pp. 915–925, 2017.

[4] A. R. Persico, P. Kirkland, C. Clemente, J. J. Soraghan, and M. Vasile, "CubeSat-based passive bistatic radar for space situational awareness: a feasibility study," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 55, no. 1, pp. 476–485, 2019.

[5] C. W. Wang, J. F. Pei, R. F. Wang, Y. L. Huang, and J. Y. Yang, "ExoMars spacecraft detection with European space surveillance bistatic radar," in *Proceedings of the 2016 CIE International Conference on Radar (RADAR)*, pp. 10–13, Guangzhou, China, October 2016.

[6] L. Mazal and P. Gurfil, "Cluster flight algorithms for disaggregated satellites," *Journal of Guidance, Control, and Dynamics*, vol. 36, no. 1, pp. 124–135, 2013.

[7] A. Kandhalu and R. Rajkumar, "QoS-based resource allocation for next-generation spacecraft networks," in *Proceedings of the IEEE 33rd Real Time Systems Symposium*, pp. 163–172, San Juan, Puerto Rico, December 2012.

[8] S. Nag, C. K. Gatebe, and O. d. Weck, "Observing system simulations for small satellite formations estimating bidirectional reflectance," *International Journal of Applied Earth Observation and Geoinformation*, vol. 43, pp. 102–118, 2015.

[9] H. Zhang and P. Gurfil, "Distributed control for satellite cluster flight under different communication topologies," *Journal of Guidance, Control, and Dynamics*, vol. 39, pp. 617–627, 2015.

[10] T. Yan, S. Hu, and J. Mo, "Path formation time in the noise-limited fractionated spacecraft network with FDMA," *International Journal of Aerospace Engineering*, vol. 2018, Article ID 9124132, , 2018.

[11] J. Mo, S. Hu, T. Yan, X. Song, and Y. Shi, "Transmit power allocation with connectivity probability for Multi-QoS in cluster flight spacecraft network," *Wireless Communications and Mobile Computing*, vol. 2020, Article ID 8676835, , 2020.

[12] J. Mo, S. Hu, Y. Shi, X. Song, and T. Yan, "Nodal distance distributions in cluster flight spacecraft network," *Mathematical Methods in the Applied Sciences*, vol. 43, no. 17, 2020.

[13] S. Cakaj, B. Kamo, and A. Rakipi, "The coverage analysis for low earth orbiting satellites at low elevation," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 6, 2014.

[14] P. Zong and S. Kohani, "Optimal satellite LEO constellation design based on global coverage in one revisit time," *International Journal of Aerospace Engineering*, vol. 2019, Article ID 4373749, , 2019.

[15] W. Beide, "The nature of bistatic and multistatic radar," in *Proceedings of the 2001 CIE International Conference on Radar*, pp. 882–884, Edinburgh, UK, 2001.

[16] S. D. Blunt and K. Gerlach, "Aspects of multistatic adaptive pulse compression," in *Proceedings of the International IEEE Radar Conference*, pp. 104–108, Washington, DC, USA, June 2005.

[17] I. Bradaric, G. Capraro, D. D. Weiner, and M. C. Wicks, "Multistatic radar systems signal processing," in *Proceedings of the IEEE Radar Conference*, pp. 106–113, New Delhi, India, May 2006.

[18] D. Bruyère and N. A. Goodman, "Performance of multistatic space-time adaptive processing," in *Proceedings of the IEEE Radar Conference*, pp. 533–538, Verona, NY, USA, 2006.

[19] E. Hanle, "Survey of bistatic and multistatic radar," *IEE Proceedings F Communications, Radar and Signal Processing*, vol. 133, no. 7, pp. 587–595, 1986.

[20] M.-Z. Shieh, S.-C. Tsai, and M.-C. Yang, "On the inapproximability of maximum intersection problems," *Information Processing Letters*, vol. 112, no. 19, pp. 723–727, 2012.

[21] E. T. Bogue, C. C. Souza, E. C. Xavier, and A. S. Freire, "An integer programming formulation for the maximum *k*-subset intersection problem," *Lecture Notes in Computer Science*, vol. 8596, pp. 87–99, 2014.

[22] X. W. Gong, J. S. Zhang, C. Cochran, and K. Xing, "Optimal placement for barrier coverage in bistatic radar sensor networks," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 259–271, 2014.

[23] B. Wang, J. Y. Chen, W. Y. Liu, and L. T. Yang, "Minimum cost placement of bistatic radar sensors for belt barrier coverage," *IEEE Transactions on Computers*, vol. 65, no. 2, pp. 577–588, 2015.

[24] Q. Q. Yang, S. B. He, and J. M. Chen, "Energy-efficient area coverage in bistatic radar sensor networks," in *Proceedings of the Global Communications Conference*, pp. 280–285, Atlanta, GA, USA, December 2013.

[25] X. W. Gong, J. S. Zhang, and D. A. Cochran, "A coverage theory of bistatic radar networks: worst-case intrusion path and optimal deployment," *Networking and Internet Architecture*, http://arxiv.org/abs/1206.1355, 2012.

[26] R. Q. Wang, S. B. He, J. M. Chen, Z. G. Shi, and F. Hou, "Energy-efficient barrier coverage in bistatic radar sensor networks," in *Proceedings of the IEEE International Conference on Communications (ICC)*, pp. 8–12, London, UK, 2015.

[27] J. He, Z. Xing, R. Hu et al., "Directional antenna intelligent coverage method based on traversal optimization algorithm," *Computers, Materials & Continua*, vol. 60, no. 2, pp. 527–544, 2019.

[28] C. Duan, J. Feng, H. Chang, J. Pan, and L. Duan, "Research on sensor network coverage enhancement based on non-cooperative games," *Computers, Materials & Continua*, vol. 60, no. 3, pp. 989–1002, 2019.

[29] C. Chen and Y. Lin, "Development of a data-driven ANFIS model by using PSO-LSE method for nonlinear system identification," *Intelligent Automation and Soft Computing*, vol. 25, no. 2, pp. 319–327, 2019.

[30] W. Zhang, W. Shi, and B. Sun, "BDI agent and QPSO-based parameter optimization for a marine generator excitation controller," *Intelligent Automation and Soft Computing*, vol. 25, no. 3, pp. 423–431, 2019.

[31] J. Wang, Y. Gao, C. Zhou, R. Simon Sherratt, and L. Wang, "Optimal coverage multi-path scheduling scheme with multiple mobile sinks for WSNs," *Computers, Materials & Continua*, vol. 62, no. 2, pp. 695–711, 2020.

[32] N. Pierdicca, L. De Titta, L. Pulvirenti, and G. Della Pietra, "Bistatic radar configuration for soil moisture retrieval: analysis of the spatial coverage," *Sensors*, vol. 9, no. 9, pp. 7250–7265, 2009.

[33] M. D. Errico and A. Moccia, "Remote sensing satellite formation for bistatic synthetic aperture radar observation," *Sensors, Systems, and Next-Generation Satellites V*, vol. 4540, 2001.

[34] R. R. Scherberger, H. Kaess, S. Brückner, and M. G. Gao, "Studies on the action of an anticholinergic agent in combination with a tranquilizer on gastric juice secretion in man," in *Proceedings of the 2008 9th International Conference on Signal Processing*, pp. 1460–1463, October 2008.

[35] J. H. Liu and G. H. Liao, "Spaceborne-airborne bistatic radar clutter modeling and analysis," in *Proceedings of the 2011 IEEE CIE International Conference on Radar*, pp. 24–27, Chengdu, China, October 2011.

[36] J. Huang, Y. Su, L. Huang, W. Liu, and F. Wang, "An optimized snapshot division strategy for satellite network in GNSS," *IEEE Communications Letters*, vol. 20, no. 12, pp. 2406–2409, 2016.

[37] N. J. Willis and H. D. Griffifiths, "Advances in bistatic radar (book review)," *IEEE Aerospace and Electronic Systems Magazine*, vol. 23, no. 7, p. 46, 2008.

[38] L. E. Miller, "Distribution of link distances in a wireless network," *Journal of Research of the National Institute of Standards and Technology*, vol. 106, no. 2, pp. 401–412, 2001.

[39] C. C. Tseng, H. T. Chen, and K. C. Chen, "On the distance distributions of the wireless ad hoc networks," in *Proceedings of the 2006 IEEE 63rd Vehicular Technology Conference*, pp. 772–776, Melbourne, Australia, 2006.

[40] D. Patel and M. J. Jha, "Hungarian method based resource scheduling algorithm in cloud computing," *International Journal of Advance Research and Innovative Ideas in Education*, vol. 2, pp. 54–59, 2016.

[41] H. Nakayama, "Multi-objective optimization and its engineering applications," *Dagstuhl Seminar Proceedings*, vol. 2005, Article ID 04461, 2005.

[42] I. Ibrahim, Z. M. Yusof, S. W. Nawawi et al., "A novel multi-state particle swarm optimization for discrete combinatorial optimization problems," in *Proceedings of the 2012 Fourth International Conference on Computational Intelligence, Modelling and Simulation*, pp. 25–27, Kuantan, Malaysia, September 2012.

[43] M. Rosendo and A. Pozo, "Applying a discrete particle swarm optimization algorithm to combinatorial problems," in *Proceedings of the 2010 Eleventh Brazilian Symposium on Neural Networks*, pp. 23–28, Paulo, Brazil, October 2010.

[44] X. Yu, W.-N. Chen, T. Gu et al., "Set-based discrete particle swarm optimization based on decomposition for permutation-based multiobjective combinatorial optimization problems," *IEEE Transactions on Cybernetics*, vol. 48, no. 7, pp. 2139–2153, 2018.

[45] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, pp. 1942–1948, Perth, Australia, 1995.

[46] Y. Yang, W. Yi, T. Zhang et al., "Fast optimal antenna placement for distributed MIMO radar with surveillance performance," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1955–1959, 2015.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 8877037, 12 pages
https://doi.org/10.1155/2021/8877037

Research Article

An Efficient Polynomial Time Algorithm for a Class of Generalized Linear Multiplicative Programs with Positive Exponents

Bo Zhang,[1] YueLin Gao,[2,3] Xia Liu,[1] and XiaoLi Huang[3]

[1]*School of Mathematics and Statistics, Ningxia University, Yinchuan 750021, China*
[2]*Ningxia Province Key Laboratory of Intelligent Information and Data Processing, North Minzu University, Yinchuan 750021, China*
[3]*School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, China*

Correspondence should be addressed to YueLin Gao; gaoyuelin@263.net

Received 23 September 2020; Accepted 6 February 2021; Published 26 February 2021

Academic Editor: Guoqiang Wang

This paper explains a region-division-linearization algorithm for solving a class of generalized linear multiplicative programs (GLMPs) with positive exponent. In this algorithm, the original nonconvex problem GLMP is transformed into a series of linear programming problems by dividing the outer space of the problem GLMP into finite polynomial rectangles. A new two-stage acceleration technique is put in place to improve the computational efficiency of the algorithm, which removes part of the region of the optimal solution without problems GLMP in outer space. In addition, the global convergence of the algorithm is discussed, and the computational complexity of the algorithm is investigated. It demonstrates that the algorithm is a complete polynomial time approximation scheme. Finally, the numerical results show that the algorithm is effective and feasible.

1. Introduction

Consider a class of generalized linear multiplicative programs (GLMPs):

$$\text{(LFP)}: \begin{cases} \min \quad f(x) = \prod_{i=1}^{p} \left(c_i^T x + d_i\right)^{\alpha_i} \\ \text{s.t. } x \in X = \{x \in \mathbb{R}^n | Ax \le b, x \ge 0\}. \end{cases} \tag{1}$$

Here, $p \ge 2$, X is a nonempty bounded closed set, $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, c_i \in \mathbb{R}^n$, $d_i \in \mathbb{R}$, and $\alpha_i > 0$. T represents the transpose of a vector (e.g., c_i^T represents the transpose of a vector c_i). Besides, we assume that for any $x \in$ X, all make $c_i^T x + d_i > 0, i = 1, 2, \ldots, p$.

The problem GLMP usually has multiple nonglobal local optimal solutions and is a class of NP-hard problems [1], which can be widely used in the fields of finance optimization [2, 3], robust optimization [4], microeconomics [5], and multiobjective decision making [6, 7]. In addition, the GLMP also includes a wide range of mathematical programming categories, such as linear multiplicative programming, quadratic programming, bilinear programming, and so on. Therefore, for these and various other reasons, GLMP has caught the attention of many experts, scholars, and engineering practitioners who have studied this theory and set off a new wave of global optimization learning. With the increasing dependence of practical problems on modeling optimization, local optimization theory and global optimization algorithms have made remarkable progress. However, compared with local optimization algorithm, the theory of global optimization algorithm is still quite insufficient. There are many methods to study this kind of problems, such as level set algorithm [8], heuristic algorithm [9, 10], branch and bound algorithm [11–13], outer approximation algorithm [14], parametric simplex algorithm [15], and so on, but these methods do not give the computational complexity of the algorithm. In addition, Depetrini and Locatelli [16] considered the problem of minimizing the product of two affine functions over a polyhedron set and proposed a polynomial time approximation algorithm. Locatelli [17] presented an approximate algorithm for solving more general types of global

optimization problems and deduced the computational complexity of the algorithm, but the numerical results of the algorithm are lacking. Recently, Shen and Wang [18] also proposed a full polynomial time approximation algorithm for resolving the problem GLMP globally, but there is no acceleration technique. Moreover, for a more comprehensive overview of the GLMP, we encourage the readers to go through the more detailed literature [8, 19–21].

In this paper, in order to solve the GLMP, two approximation algorithms are proposed, which is mainly by establishing a nonuniform grid; the process of solving the original problem is transformed into the process of solving a series of linear problems; it is proved that the proposed algorithm can obtain a global ε–approximation solution for GLMP. Besides, we put forward a two-stage acceleration technique to speed up Algorithm 1, which yields Algorithm 2. Then, by discussing the computational complexity of the algorithm, it is shown that the two algorithms are polynomial time approximation algorithms. Numerical experiments show that the performance of Algorithm 2 is obviously better than that of Algorithm 1, and the numerical results in Table 2 show that in solving problem 1-3, Algorithm 2 uses less CPU running time and iterations than [17, 18].

The rest of this paper is organized as follows. In Section 2, we first transform the problem GLMP into its equivalent optimization problem EOP and give its region-decomposition-linearization technique. Section 3 presents the global ε–approximation algorithm for problem GLMP and obtains the convergence for the proposed algorithm. In Section 4, we give the computational complexity for the proposed algorithm and carry out some numerical experiments in Section 5 to verify the feasibility and effectiveness of the algorithm. The concluding section is a simple summary.

2. Equivalence Problem and Its Linearization Technique

In this section, we will give the equivalent optimization problem EOP of the problem GLMP, then give the corresponding properties by studying the objective function of the EOP, and then explain the linearization technique of the equivalent problem.

2.1. Equivalent Problems and Their Properties. In order to solve the problem GLMP, the definition of global ϵ–approximation solution is given below.

Definition 1. Let x^* be a global optimal solution to the problem GLMP at a given precision $\varepsilon \in (0,1)$. If $\hat{x} \in X$ satisfies $f(\hat{x}) \le (1+\varepsilon) f(x^*)$, $\hat{x}$ is referred to as the global approximation of the problem GLMP.

To obtain the global ε–approximation solution for GLMP, let $f_i(x) = c_i^T x + d_i$, $l_i = \min_{x\in X} f_i(x)$.

Theorem 1. *For each $i = 1,2,\ldots,p$, let $\tilde{x}^i = \operatorname{argmin}_{x\in X} f_i(x)$, $Q = \cup_{i=1}^p \tilde{x}^i$, $\check{x} = \operatorname{argmin}_{x\in Q} f(x)$, $\tilde{U} = f(\check{x})$. Then, for each $i \in \{1,2,\ldots,p\}$, let $M_i = \prod_{j=1, j\ne i}^p l_j^{\alpha_j}$; then, $f_i(x^*) \le u_i$ with $u_i = (\tilde{U}/M_i)^{(1/\alpha_i)}$.*

Proof. It is easy to know that for any $i \in \{1,2,\ldots,p\}$, there are $l_i \le f_i(x^*)$; thus,

$$\prod_{j=1, j\ne i}^{p} l_j^{\alpha_j} (f_i(x^*))^{\alpha_i} \le \prod_{i=1}^{p} (f_i(x^*))^{\alpha_i} = f(x^*) \le f(\check{x}) = \tilde{U}. \tag{2}$$

Therefore, $f_i(x^*) \le (\tilde{U}/M_i)^{(1/\alpha_i)} = u_i$ and then the conclusion holds.

Next, according to Theorem 1, for each $i = 1,2,\ldots,p$, $u_i = (\tilde{U}/M_i)^{(1/\alpha_i)}$ provide an upper bound for every $f_i(x^*)$.

On the basis of the above definition of l_i and u_i, define the rectangle H as follows.

$$H = [l_1, u_1] \times [l_2, u_2] \times \cdots \times [l_p, u_p]. \tag{3}$$

Moreover, the rectangle H is also called the outer space of the GLMP. Thus, by introducing variable $y = (y_1, y_2, \ldots, y_p)^T \in H$, the problem GLMP is equivalent to the following problem P1.

$$(\text{P1}) \begin{cases} \min & h(y) = \prod_{i=1}^{p} y_i^{\alpha_i}, \\ \text{s.t.} & f_i(x) \le y_i, i = 1,2,\ldots,p \\ & x \in X, y \in H. \end{cases} \tag{4}$$

Next, the equivalence of problems GLMP and P1 is explained by Theorem 1. □

Theorem 2. *x^* is a global optimal solution of problem GLMP if and only if (x^*, y^*) is an optimal solution of problem P1 and $y_i^* = f_i(x^*), i = 1,2,\ldots,p$.*

Proof. Let $y_i^* = f_i(x^*), i = 1,2,\ldots,p$ if x^* is a global optimal solution of the problem GLMP. Then, then it is obvious that (x^*, y^*) is a feasible solution to P1. Suppose (x^*, y^*) is not an optimal solution of P1; then, there is at least one feasible solution $(\overline{x}, \overline{y})$ of P1, which makes

$$f(\overline{x}) = \prod_{i=1}^{p} (f_i(\overline{x}))^{\alpha_i} \le \prod_{i=1}^{p} \overline{y}_i^{\alpha_i} < \prod_{i=1}^{p} (y_i^*)^{\alpha_i} = \prod_{i=1}^{p} (f_i(x^*))^{\alpha_i} = f(x^*), \tag{5}$$

which contradicts the optimality of the x^*, so the hypothesis does not hold, and then (x^*, y^*) is an optimal solution of P1.

Conversely, if (x^*, y^*) is an optimal solution for P1 and if there is a $i \in \{1,2,\ldots,p\}$ that makes $f_i(x^*) < y_i^*$, let $\tilde{y}_i = f_i(x^*)$, then $(x^*, \tilde{y})$ is a feasible solution for P1 and

$$\prod_{i=1}^{p} \tilde{y}_i^{\alpha_i} < \prod_{i=1}^{p} (y_i^*)^{\alpha_i}, \tag{6}$$

which contradicts the optimality of (x^*, y^*), so $y_i^* = f_i(x^*), i = 1,2,\ldots,p$. Suppose x^* is not a global optimal solution of the problem GLMP; then, there must be a $\overline{x} \in X$ that makes $f(\overline{x}) < f(x^*)$. Let $\overline{y}_i = f_i(\overline{x})$; obviously, $(\overline{x}, \overline{y})$ is a feasible solution to P1, so we have

(1) **Step 0 (initialization)**. Set $\varepsilon \in (0,1), \delta = (1+\varepsilon)^{(1/\rho)}, F = +\infty, k = 0$. By using formulas (22) and (23), the ratio used for the two consecutive segments in each dimension is δ, which subdivides H into smaller rectangles. Represent the vertex of each small rectangle as $\nu = (\nu_1, \nu_2, \ldots, \nu_p)$, which is stored in the set B^δ.
(2) **Step 1**. Select a point ν from the B^δ, solve the linear programming problem (LP_ν), and let $B^\delta = B^\delta \backslash \nu$.
(3) **Step 2**. If the problem (LP_ν) is solvable, then $D(\nu) \neq \varnothing$, and let $g(\nu) = \prod_{i=1}^{p} (\nu_i)^{\alpha_i}$; if $g(\nu) < F$, let $F = g(\nu)$, $\overline{\nu} = \nu$, $x_{\overline{\nu}} = x_\nu$; if $B^\delta \neq \varnothing$, set $k = k+1$ and go to Step 1; otherwise, the algorithm terminates; let

Algorithm 1: Original algorithm.

$$\prod_{i=1}^{p} (\overline{y}_i)^{\alpha_i} = f(\overline{x}) < f(x^*) = \prod_{i=1}^{p} (y_i^*)^{\alpha_i}, \tag{7}$$

which contradicts the optimality of (x^*, y^*). Therefore, x^* is the global optimal solution of the problem GLMP, which proves to be completed.

It is easy to understand from Theorem 2 that the problems GLMP and P1 are equivalent and have the same global optimal value.

Then, for a given $y \in H$, define the set

$$D(y) = \{x \in X \mid f_i(x) \leq y_i,\ i = 1, 2, \ldots, p\}, \tag{8}$$

and function

$$g(y) = \begin{cases} h(y), & D(y) \neq \varnothing, \\ +\infty, & D(y) = \varnothing. \end{cases} \tag{9}$$

Then, the problem P1 is equivalent to the following equivalent optimization problem.

$$(\mathrm{EOP}) \begin{cases} \min & g(y) \\ \text{s.t.} & y \in H. \end{cases} \tag{10}$$

□

Theorem 3. *y^* is the global optimal solution of the problem EOP if and only if (x^*, y^*) is the optimal solution of P1 and $y_i^* = f_i(x^*), i = 1, 2, \ldots, p$.*

Proof. Suppose (x^*, y^*) is an optimal solution of P1; then, according to Theorem 2, we can know $y_i^* = f_i(x^*), i = 1, 2, \ldots, p$ and $y^* \in H$. In addition, $h(y^*) = g(y^*) = \prod_{i=1}^{p} (y_i^*)^{\alpha_i}$. Suppose that y^* is not the global optimal solution of the problem EOP; there must be a $\overline{y} \in H$ such that $g(\overline{y}) < g(y^*)$ and $D(\overline{y}) \neq \varnothing$; then, there must also be a $\overline{x} \in D(\overline{y})$ such that $f_i(\overline{x}) \leq \overline{y}_i, i = 1, 2, \ldots, p$. Then, $(\overline{x}, \overline{y})$ is a feasible solution of P1; there is $h(\overline{y}) = g(\overline{y}) < g(y^*) = h(y^*)$, which contradicts the optimality of (x^*, y^*), so the hypothesis does not hold, so y^* is the global optimal solution of the problem.

On the other hand, if y^* is a global optimal solution of the problem EOP, then $D(y^*) \neq \varnothing$, and there must be a $x^* \in D(y^*)$ such that (x^*, y^*) is a feasible solution of P1. Suppose (x^*, y^*) is not the global optimal solution of the problem P1; then, there must be an optimal solution $(\overline{x}, \overline{y})$ to the problem P1 such that $h(\overline{y}) < h(y^*), \overline{y}_i = f_i(\overline{x}), i = 1, 2, \ldots, p$, so $D(\overline{y}) \neq \varnothing$ and $g(\overline{y}) = h(\overline{y}) < h(y^*) = g(y^*)$, which contradicts the fact that y^* is the global optimal solution of the problem EOP. Therefore, (x^*, y^*) is the global optimal solution of P1, and $y_i^* = f_i(x^*), i = 1, 2, \ldots, p$ can be obtained from Theorem 2 and then proved to be over.

Through Theorem 3, the problems EOP and P1 have the same global optimal value, so combined with Theorem 2, the problems EOP and GLMP are also equivalent. Therefore, we can solve the equivalent problem EOP instead of addressing the problem GLMP.

Next, we consider the following linear programming problem:

$$\mathrm{LP}_y \begin{cases} \min & \displaystyle\sum_{i=1}^{p} \frac{\alpha_i f_i(x)}{y_i} \\ \text{s.t.} & x \in D(y). \end{cases} \tag{11}$$

If $D(y) \neq \varnothing$, the optimal solution to the problem LP_y is recorded as x_y, and let $\widetilde{y}_i = f_i(x_y)$, $\rho = \sum_{i=1}^{p} \alpha_i > 0$; then,

$$\rho = \sum_{i=1}^{p} \frac{\alpha_i y_i}{y_i} \geq \sum_{i=1}^{p} \frac{\alpha_i f_i(x)}{y_i}, \quad \forall x \in D(y). \tag{12}$$

Furthermore, according to the Jensen inequality, we have

$$\sum_{i=1}^{p} \frac{\alpha_i f_i(x_y)}{y_i} \geq \rho \left(\prod_{i=1}^{p} \left(\frac{f_i(x_y)}{y_i} \right)^{\alpha_i} \right)^{(1/\rho)} = \rho \left(\frac{g(\widetilde{y})}{g(y)} \right)^{(1/\rho)}, \tag{13}$$

and then

$$\rho \geq \rho \left(\frac{g(\widetilde{y})}{g(y)} \right)^{(1/\rho)}, \quad g(\widetilde{y}) \leq g(y). \tag{14}$$

□

Theorem 4. *Suppose $x^* \in X$ is a global optimal solution of the original problem GLMP; let $y_i^* = f_i(x^*), i = 1, 2, \ldots, p$; then, $y^* = (y_1^*, y_2^*, \ldots, y_p^*)^T \in H$ and x^* is also a global optimal solution of the problem (LP_{y^*}).*

Proof. Firstly, according to Theorems 2 and 3, we know that y^* is a global optimal solution of the problem EOP. Then, by using formula (14) and the optimality of the global optimal solution y^* of the EOP, we can see that x^* is an optimal solution of the problem (LP_{y^*}).

Next, the properties of the function $g(y)$ over H are given by Theorem 5. □

Theorem 5. *For a given precision $\varepsilon \in (0,1)$, let $\delta = (1+\varepsilon)^{(1/\rho)}$; then, for any $\overline{y} \in H$, there is*

$$g(\overline{y}) \le (1+\varepsilon) g(y), \quad \forall y \in \left[\frac{\overline{y}}{\delta}, \overline{y}\right]. \tag{15}$$

In addition, if $D(\overline{y}) \neq \varnothing$, the optimal solution to the problem $(\mathrm{LP}_{\overline{y}})$ is recorded as $\overline{x}$; then, let $\widetilde{y}_i = f_i(\overline{x})\,(i = 1, 2, \ldots, p)$; there is also

$$g(\widetilde{y}) \le g(\overline{y}) \le (1+\varepsilon) g(y), \quad \forall y \in \left[\frac{\overline{y}}{\delta}, \overline{y}\right]. \tag{16}$$

Proof. For all $\overline{y} \in H$, according to the definition of $D(y)$ and $\delta = (1+\varepsilon)^{(1/\rho)} > 1$, one can know $D(\overline{y}/\delta) \subseteq D(\overline{y})$.

If $D(\overline{y}/\delta) \neq \varnothing$, for any $y \in [(\overline{y}/\delta), \overline{y}]$, we have $D(y) \neq \varnothing$; obviously, $g(\widetilde{y}) \le g(\overline{y})$ and $y_i \ge (\overline{y}_i/\delta)$ for each $i = 1, 2, \ldots, p$. Thus,

$$\prod_{i=1}^{p} \left(\frac{\overline{y}_i}{\delta}\right)^{\alpha_i} \le \prod_{i=1}^{p} y_i^{\alpha_i}. \tag{17}$$

Moreover, according to the definition of function $g(y)$, $g(y) = \prod_{i=1}^{p} y_i^{\alpha_i}$; thus,

$$g\left(\frac{\overline{y}}{\delta}\right) = \prod_{i=1}^{p} \left(\frac{\overline{y}_i}{\delta}\right)^{\alpha_i} = \frac{1}{\delta^{\rho}} \prod_{i=1}^{p} \overline{y}_i^{\alpha_i} = \frac{1}{\delta^{\rho}} g(\overline{y}). \tag{18}$$

And in combination with the formulas (17) and (18), we have

$$g(y) \ge g\left(\frac{\overline{y}}{\delta}\right) = \frac{1}{\delta^{\rho}} g(\overline{y}), \quad \forall y \in \left[\frac{\overline{y}}{\delta}, \overline{y}\right]. \tag{19}$$

Further, through formula (19) and combined with the definition of δ, we can understand that formula (16) is formed, and formula (15) is of course also true.

If $D(\overline{y}/\delta) = \varnothing, D(\overline{y}) \neq \varnothing$, it is clear that the inequality $g(\widetilde{y}) \le g(\overline{y})$ is established.

For all $y \in [(\overline{y}/\delta), \overline{y}]$, if $D(y) \neq \varnothing$, we have $y_i \ge (\overline{y}_i/\delta)\,(i = 1, 2, \ldots, p)$, and $y \neq (\overline{y}/\delta)$; then,

$$\prod_{i=1}^{p} \left(\frac{\overline{y}_i}{\delta}\right)^{\alpha_i} \le g(y) = \prod_{i=1}^{p} y_i^{\alpha_i}. \tag{20}$$

Besides,

$$g(\overline{y}) = \prod_{i=1}^{p} \overline{y}_i^{\alpha_i} = \delta^{\rho} \prod_{i=1}^{p} \left(\frac{\overline{y}_i}{\delta}\right)^{\alpha_i}. \tag{21}$$

By using the definition of δ and formulas (20) and (21), one can infer that formulas (15) and (16) hold.

If $D(y) = \varnothing$ and $g(y) = +\infty$, then formulas (15) and (16) obviously hold.

If $D(\overline{y}) = \varnothing$, the problem $(\mathrm{LP}_{\overline{y}})$ is not solved, and for any $y \in [(\overline{y}/\delta), \overline{y}]$, there is $D(y) = \varnothing$, then $g(y) = +\infty$, so formula (15) is clearly established and the proof of the conclusion is completed.

Theorem 5 shows that for any $\overline{y} \in H$, we can determine whether the $D(\overline{y})$ is not empty by solving the linear programming problem $(\mathrm{LP}_{\overline{y}})$ and then determine whether formula (16) holds. □

2.2. Linearization Techniques. The objective function of the problem EOP is still nonconvex compared to the problem GLMP. But the space H in which the variable y of the objective function is located is p dimensions. Therefore, based on the above discussion, in order to solve the EOP, for a given $\varepsilon \in (0, 1)$, we first split the outer space H on each dimension at a ratio of $\delta = (1+\varepsilon)^{(1/\rho)}$, thus producing several small rectangles.

To do this, let

$$\gamma_i = \arg\max\{\sigma \in \mathbb{N} | l_i \delta^{\sigma} \le u_i\}, \quad i = 1, 2, \ldots, p, \tag{22}$$

where $\mathbb{N}$ represents a non-negative integer set. Therefore, the number of these small rectangles is finite, and the set of all their vertices is

$$B^{\delta} = \left\{\nu_1, \nu_2, \ldots, \nu_p | \nu_i \in P_i^{\delta}, i = 1, 2, \ldots, p\right\}, \tag{23}$$

where $P_i^{\delta} = \{l_i, l_i\delta, \ldots, l_i\delta^{\gamma_i}\}$. Obviously, for each $y \in H$, there must be a vertex $(\nu_1, \nu_2, \ldots, \nu_p) \in B^{\delta}$ making $y_i \in [\nu_i, \delta\nu_i], i = 1, 2, \ldots, p$. Then, it can be concluded that the rectangle H can be approximated by the set B^{δ}.

Next, by using the set B^{δ}, the process of solving the problem EOP can be transformed into solving a series of subproblems. To this end, for each $\nu \in B^{\delta}$, we need to consider the value of the $g(\nu)$, that is, we need to determine whether the set $D(\nu)$ is not empty. According to Theorem 5, we can determine whether $D(\nu)$ is not empty by solving the linear programming problem (LP_{ν}). Therefore, for each vertex $\nu \in B^{\delta}$, the following linear programming subproblem needs to be solved here, that is,

$$(\mathrm{LP}_{\nu}) \begin{cases} \min \quad \displaystyle\sum_{i=1}^{p} \frac{\alpha_i f_i(x)}{\nu_i} \\ \text{s.t.} \quad \begin{array}{l} f_i(x) \le \nu_i, i = 1, 2, \ldots, p, \\ x \in X. \end{array} \end{cases} \tag{24}$$

On the basis of the conclusion of Theorem 5, if the problem (LP_{ν}) can be solved (its solution is recorded as x_{ν}), then

$$\widetilde{\nu} = \left(f_1(x_{\nu}), f_2(x_{\nu}), \ldots, f_p(x_{\nu})\right)^T \in H, \tag{25}$$

and thus

$$g(\widetilde{\nu}) \le g(\nu) \le (1+\varepsilon) g(y), \quad \forall y \in \left[\frac{\nu}{\delta}, \nu\right]. \tag{26}$$

3. Analysis of Algorithm and Its Computational Complexity

This section brings an approximate algorithm based on linearization-decomposition to solve the problem EOP. After that, the analysis of its computational complexity is proved accordingly.

3.1. Approximate Algorithm. To solve the EOP, we subdivide the external space H into a finite number of small rectangles with ratio δ and put all the vertices of these small rectangles into the set B^{δ}.

Then, for each vertex $\nu \in B^{\delta}$, by solving the linear programming problem (LP_{ν}), if (LP_{ν}) is feasible and has an optimal solution x_{ν}, then $D(\nu) \neq \varnothing$, and we can obtain a feasible solution $\tilde{\nu}$ (formula (25)) of the EOP according to x_{ν}, which makes

$$g(\tilde{\nu}) \leq g(\nu) \leq (1+\varepsilon) g(y), \quad \forall y \in \left[\frac{\nu}{\delta}, \nu\right]. \tag{27}$$

If there is a $\tilde{\nu}$ that satisfies $g(\tilde{\nu}) \leq (1+\varepsilon) g(y^{*})$, then

$$f(x_{\nu}) = \prod_{i=1}^{p} \left(f_i(x_{\nu})\right)^{\alpha_i} = \prod_{i=1}^{p} \tilde{\nu}_i^{\alpha_i} = g(\tilde{\nu}) \leq (1+\varepsilon) g(y^{*}) = (1+\varepsilon) f(x^{*}), \tag{28}$$

and thus x_{ν} is a global ε–approximation solution of the problem GLMP. The specific algorithm steps are as follows.

(1) **Step 0 (initialization).** Set $\varepsilon \in (0,1), \delta = (1+\varepsilon)^{(1/\rho)}, F = +\infty,\ k = 0$. By using formulas (22) and (23), the ratio used for the two consecutive segments in each dimension is δ, which subdivides H into smaller rectangles. Represent the vertex of each small rectangle as $\nu = (\nu_1, \nu_2, \ldots, \nu_p)$, which is stored in the set B^{δ}.

(2) **Step 1.** Select a point ν from the B^{δ}, solve the linear programming problem (LP_{ν}), and let $B^{\delta} = B^{\delta} \backslash \nu$.

(3) **Step 2.** If the problem (LP_{ν}) is solvable, then $D(\nu) \neq \varnothing$, and let $g(\nu) = \prod_{i=1}^{p} (\nu_i)^{\alpha_i}$; if $g(\nu) < F$, let $F = g(\nu)$, $\bar{\nu} = \nu$, $x_{\bar{\nu}} = x_{\nu}$; if $B^{\delta} \neq \varnothing$, set $k = k+1$ and go to Step 1; otherwise, the algorithm terminates; let

$$\tilde{\nu}_i = f_i(x_{\bar{\nu}}), \quad i = 1, 2, \ldots, p, \tilde{\nu} = \left(\tilde{\nu}_1, \tilde{\nu}_2, \ldots, \tilde{\nu}_p\right)^{T}, \tag{29}$$

and then $x_{\bar{\nu}}$, $\tilde{\nu}$ is a global ε–approximation solution to problems GLMP and EOP, respectively.

Theorem 6. *For a given precision $\varepsilon \in (0,1)$, let $\delta = (1+\varepsilon)^{(1/\rho)}$, $\bar{\nu} = \arg\min\{g(\nu) | \nu \in B^{\delta}\}$, and $x_{\bar{\nu}}$ be an optimal solution of the linear programming problem $(LP_{\bar{\nu}})$. Then, Algorithm 1 will get a global ε–approximation solution $x_{\bar{\nu}}$ for problem GLMP, i.e.,*

$$f(x_{\bar{\nu}}) \leq (1+\varepsilon) f(x^{*}), \tag{30}$$

where x^{*} is the global optimal solution to the original problem GLMP.

Proof. Let

$$y_i^{*} = f_i(x^{*}), \quad i = 1, 2, \ldots, p. \tag{31}$$

According to Theorem 1, we have

$$l_i \leq y_i^{*} \leq u_i, \quad i = 1, 2, \ldots, p. \tag{32}$$

Then, formula (32) implies that $y^{*} = (y_1^{*}, y_2^{*}, \ldots, y_p^{*})^{T} \in H$, so there must be a $\nu^{*} \in B^{\delta}$ which makes

$$\frac{\nu_i^{*}}{\delta} \leq y_i^{*} \leq \nu_i^{*}, \quad i = 1, 2, \ldots, p. \tag{33}$$

So, using Theorem 5 on the small rectangle $[(\nu^{*}/\delta), \nu^{*}]$, there will be

$$f(x^{*}) = \prod_{i=1}^{p} (y_i^{*})^{\alpha_i} = g(y^{*}) \geq \prod_{i=1}^{p} \left(\frac{\nu_i^{*}}{\delta}\right)^{\alpha_i} = \left(\frac{1}{\delta}\right)^{\sum_{i=1}^{p} \alpha_i} \prod_{i=1}^{p} (\nu_i^{*})^{\alpha_i} = \frac{1}{\delta^{\rho}} g(\nu^{*}). \tag{34}$$

Thus,

$$\delta^{\rho} f(x^{*}) = \delta^{\rho} g(y^{*}) \geq g(\nu^{*}). \tag{35}$$

Noting that $\bar{\nu} = \arg\min\{g(\nu) | \nu \in B^{\delta}\}$, we can know

$$g(\nu^{*}) \geq g(\bar{\nu}). \tag{36}$$

Since $x_{\bar{\nu}}$ is the optimal solution to the linear programming problem $(\mathrm{LP}_{\bar{\nu}})$, let

$$\tilde{\nu}_i = f_i(x_{\bar{\nu}}), \quad i = 1, 2, \ldots, p. \tag{37}$$

Apparently, $\tilde{\nu} = (\tilde{\nu}_1, \tilde{\nu}_2, \ldots, \tilde{\nu}_p) \in H$. So, by taking advantage of the formula (16) in Theorem 5, we have

$$g(\bar{\nu}) \geq g(\tilde{\nu}) = \prod_{i=1}^{p} (\tilde{\nu}_i)^{\alpha_i} = \prod_{i=1}^{p} \left(f_i(x_{\bar{\nu}})\right)^{\alpha_i} = f(x_{\bar{\nu}}). \tag{38}$$

Therefore, by integrating formulas (35) and (38) and combining the $\delta = (1+\varepsilon)^{(1/\rho)}$, we can obtain

$$f(x_{\bar{\nu}}) \leq (1+\varepsilon) f(x^{*}), \tag{39}$$

and this proof is completed. □

Remark 1. According to Theorem 6, if $y^{*} \in B^{\delta}$, then from Theorem 5, the optimal solution $x_{y^{*}}$ of the linear programming problem $(\mathrm{LP}_{y^{*}})$ is exactly the global optimal solution of the original problem GLMP.

Through Theorem 6, we can see that for a given precision $\varepsilon \in (0,1)$, Algorithm 1 will obtain a global ε–approximation solution to the problem GLMP. Moreover, Remark 1 also shows that if $y^* \in B^\delta$, then Algorithm 1 will find a global optimal solution of the problem GLMP exactly.

3.2. Accelerating Techniques. Algorithm 1 shows that, for any $\nu \in B^\delta$, it is required to solve the linear programming problem (LP_ν), in order to verify that the $D(\nu)$ is nonempty. Hence, the computational cost of Algorithm 1 depends on the number of points within the set B^δ, respectively. Then, the proposal of the acceleration technique will discard some points that are not necessary to consider the set B^δ and only consider the region that contains the global optimal solution of the problem EOP. The detailed process is given below.

If $\overline{\nu}$ is the best known solution to the problem EOP, $x_{\overline{\nu}}$ is the optimal solution to the linear programming problem $(\mathrm{LP}_{\overline{\nu}})$; for each $i = 1,2,\ldots,p$, let $\tilde{\nu}_i = f_i(x_{\overline{\nu}}), \tilde{\nu} = (\tilde{\nu}_1, \tilde{\nu}_2, \ldots, \tilde{\nu}_p)^T$; obviously $g(\tilde{\nu}) \le g(\overline{\nu})$; then, $\tilde{\nu}$ may be a better solution than $\overline{\nu}$. Well, using $\tilde{\nu}$ may be able to remove more vertices from B^δ that do not need to be explored. To give the acceleration technique for Algorithm 1, we first need to specify a necessary condition that the points in each subrectangle $H^k \subseteq H^0 = H\,(k \ge 1)$ containing the global optimal solution of the problem EOP must be satisfied, that is,

$$\prod_{i=1}^{p} l_i^{\alpha_i} \le g(y^*) \le g(y) \le g(\tilde{\nu}), \quad \forall y \in H^k, \tag{40}$$

where $H^k = [l, u^k]$, $u^k = (u_1^k, u_2^k, \ldots, u_p^k)^T$, $u_i^k \le u_i^{k-1} \le u_i$, $i = 1,2,\ldots,p$. Similarly, if $\delta = (1+\varepsilon)^{(1/p)}$ are used to segment rectangles H^k on each dimension, this will produce a limited number of small rectangles. For this purpose, let

$$\gamma_i^k = \arg\max\left\{\sigma \in \mathbb{N} \mid l_i \delta^\sigma \le u_i^k\right\}, \quad i = 1,2,\ldots,p. \tag{41}$$

Then, a set of vertices of a finite number of small rectangles will also be generated on a rectangular H^k, that is,

$$B_k^\delta = \left\{\nu_1, \nu_2, \ldots, \nu_p \mid \nu_i \in P_{ki}^\delta,\ i = 1,2,\ldots,p\right\}, \tag{42}$$

where $P_{ki}^\delta = \left\{l_i, l_i\delta, \ldots, l_i\delta^{\gamma_i^k}\right\}$. Clearly, $B_k^\delta \subseteq B_0^\delta = B^\delta$ and $B_k^\delta \subset H^k \subseteq H_0 = H$.

Based on the above discussion, we will give Propositions 1 and 2 to clarify the acceleration techniques of the algorithm.

Proposition 1. *The global optimal solution of the problem EOP cannot be obtained on the set $\overline{B}_{ki}^\delta$ if a $i \in \{1,2,\ldots,p\}$ makes $(g(\tilde{\nu})/M_i)^{(1/\alpha_i)} < l_i\delta^{\gamma_i^k}$, of which*

$$\overline{B}_{ki}^\delta = \left\{\nu \in B_k^\delta \mid \left(\frac{g(\tilde{\nu})}{M_i}\right)^{(1/\alpha_i)} < \nu_i\right\}, \quad i \in \{1,2,\ldots,p\}. \tag{43}$$

Proof. If $\nu \in \overline{B}_{ki}^\delta$, then there must be $(g(\tilde{\nu})/M_i)^{(1/\alpha_i)} < \nu_i \le l_i\delta^{\gamma_i^k}$, and thus there is

$$g(\tilde{\nu}) = \left(\left(\frac{g(\tilde{\nu})}{M_i}\right)^{(1/\alpha_i)}\right)^{\alpha_i} M_i < (\nu_i)^{\alpha_i} M_i = (\nu_i)^{\alpha_i} \prod_{j=1, j\ne i}^{p} l_j^{\alpha_j} \le \prod_{j=1}^{p} (\nu_j)^{\alpha_j} = g(\nu), \tag{44}$$

which contradicts the inequality chain (40), so the conclusion is valid.

With Proposition 1, we generate a new rectangle H^{k+1} and vertex set B_{k+1}^δ, i.e., for each $i = 1,2,\ldots,p$, let

$$u_i^{k+1} = \begin{cases} \left(\dfrac{g(\tilde{\nu})}{M_i}\right)^{(1/\alpha_i)}, & \left(\dfrac{g(\tilde{\nu})}{M_i}\right)^{(1/\alpha_i)} < l_i\delta^{\gamma_i^k}, \\ u_i^k, & \left(\dfrac{g(\tilde{\nu})}{M_i}\right)^{(1/\alpha_i)} \ge l_i\delta^{\gamma_i^k}, \end{cases} \tag{45}$$

as well as

$$\gamma_i^{k+1} = \begin{cases} \arg\max\left\{\sigma \in \mathbb{N} \mid l_i\delta^\sigma \le u_i^{k+1}\right\}, & \left(\dfrac{g(\tilde{\nu})}{M_i}\right)^{(1/\alpha_i)} < l_i\delta^{\gamma_i^k}, \\ \gamma_i^k, & \left(\dfrac{g(\tilde{\nu})}{M_i}\right)^{(1/\alpha_i)} \ge l_i\delta^{\gamma_i^k}. \end{cases} \tag{46}$$

Well, $u^{k+1} = [l, u^{k+1}]$ with $u^{k+1} = (u_1^{k+1}, u_2^{k+1}, \ldots, u_p^{k+1})$.

Moreover, the above rules may produce a small rectangular vertex set B_{k+1}^δ with relatively few new elements, but there is still $\tilde{\nu} \in B_{k+1}^\delta$, so we then give Proposition 2 to delete the other unconsidered elements in B_{k+1}^δ. □

Proposition 2. *If $\overline{\nu}$ is the best known solution to the problem EOP, $x_{\overline{\nu}}$ is the optimal solution to the linear programming problem $(LP_{\overline{\nu}})$; for each $i = 1,2,\ldots,p$, let $\tilde{\nu}_i = f_i(x_{\overline{\nu}}), \tilde{\nu} = (\tilde{\nu}_1, \tilde{\nu}_2, \ldots, \tilde{\nu}_p)^T$, and define the set*

$$\overline{B}_{k+1}^\delta = \left\{\nu \in B_{k+1}^\delta \mid \tilde{\nu}_i \le \nu_i, i = 1,2,\ldots,p\right\}. \tag{47}$$

Then, for any $\nu \in \overline{B}_{k+1}^\delta$, the EOP cannot get a better solution than $\tilde{\nu}$.

Proof. Since $x_{\overline{\nu}}$ is the optimal solution to a linear programming problem $(\mathrm{LP}_{\overline{\nu}})$, then there is at least one point $x_{\overline{\nu}}$ in the set $D(\tilde{\nu})$, so $D(\tilde{\nu}) \ne \varnothing$. For arbitrary $\nu \in \overline{B}_{k+1}^\delta$, obviously $D(\tilde{\nu}) \subseteq D(\nu)$, and thus $D(\nu) \ne \varnothing$. According to the

definition of the function $g(y)$, for each $\nu \in \overline{B}_{k+1}^{\delta}$, the objective function value of the EOP meets

$$g(\nu) = \prod_{i=1}^{p} (\nu_i)^{\alpha_i} \geq \prod_{i=1}^{p} (\tilde{\nu}_i)^{\alpha_i} = g(\tilde{\nu}), \tag{48}$$

and this conclusion is proved.

Next, for a given $\varepsilon \in (0,1)$, $\delta = (1+\varepsilon)^{(1/\rho)}$, make use of Proposition 2; let

$$\tau_i^{k+1} = \arg\min\left\{\sigma \in \mathbb{N} | \tilde{\nu}_i \leq l_i \delta^{\sigma} \leq u_i^{k+1}\right\}. \tag{49}$$

Through the expression of γ_i^{k+1} in (46), the set $\overline{B}_{k+1}^{\delta}$ is defined as follows.

$$\begin{aligned}\overline{B}_{k+1}^{\delta} = &\{l_i\delta^{\sigma_1}, l_i\delta^{\sigma_2}, \ldots, l_i\delta^{\sigma_p} | \sigma_i \\ &\in \left\{\tau_i^{k+1}, \tau_i^{k+1}+1, \ldots, \gamma_i^{k+1}\right\}, i = 1,2,\ldots,p\}.\end{aligned} \tag{50}$$

Therefore, for the convenience of narration, let $S_{k+1}^{\delta} = B_{k+1}^{\delta} \setminus \overline{B}_{k+1}^{\delta}$. This means that in order to obtain a global ε–approximation solution for problem EOP, it is only necessary to calculate up to $|S_{k+1}^{\delta}|$ linear programming subproblems (LP_{ν}) to determine whether the $D(\nu)$ is not empty, which determines the function value $g(\nu)$ at each vertex $\nu \in S_{k+1}^{\delta}$. Then, by using the set S_{k+1}^{δ}, the computational efficiency of Algorithm 1 will be improved, leading to the following algorithm.

$$\begin{aligned}&\tilde{\nu}_i = f_i(x_{\tilde{\nu}}), \quad i = 1,2,\ldots,p, \\ &\tilde{\nu} = \left(\tilde{\nu}_1, \tilde{\nu}_2, \ldots, \tilde{\nu}_p\right)^T,\end{aligned} \tag{51}$$

and then $x_{\tilde{\nu}}$, $\tilde{\nu}$ is a global ε–approximation solution to the problems GLMP and EOP, respectively.

Note that the Algorithm 2 simply removes the set of vertices that do not contain a global optimal solution; therefore, it is similar to Theorem 6; Algorithm 2 will also return a global ε–approximation solution of the problem GLMP and EOP as well. □

4. Analysis of Computational Complexity of the Algorithm

We first give Lemma 1 to discuss the computational complexity of the two algorithms.

Lemma 1 (see [22]). *Let λ be the maximum of the absolute values of all the elements A, b, c_i, d_i in problem GLMP; then, each component x_j^0 of any pole x^0 of X can be expressed as $x_j^0 = (p_j/q)$, where $0 \leq p_j \leq (n\lambda)^n$, $0 < q \leq (n\lambda)^n$, $j = 1,2,\ldots,n$.*

Because for each $i = 1,2,\ldots,p$, the solution $\tilde{x}^i$ to the linear programming problem $l_i = \min_{x\in X} f_i(x)$ is the pole of X, by Lemma 1, we have $\tilde{x}_j^i = (p_j^i/q^i)$, where $0 \leq p_j^i \leq (n\lambda)^n$, $0 < q^i \leq (n\lambda)^n$, $j = 1,2,\ldots,n$. Thus, $l_i = \sum_{j=1}^{n} c_{ij}(p_j^i/q^i) + d_i$, $i = 1,2,\ldots,p$. Moreover, let

$$\begin{aligned}&\tilde{q} = \max\left\{\frac{1}{q^i} | i = 1,2,\ldots,p\right\}, \\ &\omega = \min\{l_i | i = 1,2,\ldots,p\},\end{aligned} \tag{52}$$

$$\tilde{U} = f(\check{x}) = \min_{1\leq i\leq p} f\left(\tilde{x}^i\right), \tag{53}$$

and for the sake of the following smooth description of Theorem 7, here $\check{x}$ is defined in Theorem 1.

Theorem 7. *For a given $p \geq 2$, in order to obtain a global ε–approximation solution to the problem GLMP, the upper limit of the time required for the proposed Algorithm 1 is*

$$O\left(\left(\frac{2\tilde{\alpha}\rho^2}{\varepsilon}\left[(n+1)\ln(n\lambda) - \ln\omega\right] + 1\right)^p \cdot T(m+p,n)\right), \tag{54}$$

where $\tilde{\alpha} = \max\{(1/\alpha_i) | i = 1,2,\ldots,p\}$, $\rho = \sum_{i=1}^{p} \alpha_i$, and $T(m+p,n)$ represents the upper limit of the time used to solve a linear programming problem with $m+p$ linear constraints and n variables at a time.

Proof. From the formulas (22) and (23), we can see that the maximum number of midpoint of the set B^{δ} is

$$\prod_{i=1}^{p}\left(\log_{\delta}\frac{u_i}{l_i} + 1\right). \tag{55}$$

Using the definition of $\tilde{q}, \omega$ in formula (52) and Lemma 1, we have

$$\omega \leq l_i \leq \tilde{q} n\lambda (n\lambda)^n + \lambda \leq 2\tilde{q}(n\lambda)^{n+1}, \quad i = 1,2,\ldots,p. \tag{56}$$

Furthermore, we also have

$$\tilde{U} = \prod_{i=1}^{p}\left(c_i^T\check{x} + d_i\right)^{\alpha_i} \leq \prod_{i=1}^{p}\left(2\tilde{q}(n\lambda)^{n+1}\right)^{\alpha_i} = \left(2\tilde{q}(n\lambda)^{n+1}\right)^{\sum_{i=1}^{p}\alpha_i}, \tag{57}$$

by using formula (53) and the above inequality (56). Of course, according to the definition of M_i and u_i in Theorem 1, and in conjunction with $\rho = \sum_{i=1}^{p}\alpha_i$, there will be

$$u_i = \left(\frac{\hat{U}}{M_i}\right)^{(1/\alpha_i)} \leq \left(2\tilde{q}(n\lambda)^{n+1}\right)\left(\frac{2\tilde{q}(n\lambda)^{n+1}}{\omega}\right)^{(\rho/\alpha_i)-1}. \tag{58}$$

By means of above formulas (56) and (58), we can have

$$\frac{u_i}{l_i} \leq \left(\frac{2\tilde{q}(n\lambda)^{n+1}}{\omega}\right)^{(\rho/\alpha_i)}, \tag{59}$$

and thus

$$\begin{aligned}\ln\frac{u_i}{l_i} &\leq \frac{\rho}{\alpha_i}\left[\ln 2\tilde{q} + (n+1)\ln(n\lambda) - \ln\omega\right] \\ &\leq \rho\tilde{\alpha}\left[\ln 2\tilde{q} + (n+1)\ln(n\lambda) - \ln\omega\right].\end{aligned} \tag{60}$$

(1) **Step 0 (initialization).** Set $\varepsilon \in (0,1), \delta = (1+\varepsilon)^{(1/\rho)}$. By using formulas (22) and (23), $H^0 = H$ is subdivided into smaller rectangles, such that the ratio of two consecutive segments is δ in each dimension. Represent the vertex of each small rectangle as $\nu = (\nu_1, \nu_2, \ldots, \nu_p)$, which is stored in the set B^δ. Let $F = +\infty$, $T = \varnothing$, $B_0^\delta = B^\delta$, $\Xi^0 = B_0^\delta$, $k = 0$.
(2) **Step 1.** Select a point $\nu = (\nu_1, \nu_2, \ldots, \nu_p)^T$ from the Ξ^k, solve the linear programming problem (LP_ν), and let $T = T \cup \nu$.
(3) **Step 2.** If the problem (LP_ν) is solvable, then $D(\nu) \neq \varnothing$, and let $g(\nu) = \prod_{i=1}^{p} (\nu_i)^{\alpha_i}$; if $g(\nu) < F$, let $\overline{\nu} = \nu$, $x_{\overline{\nu}} = x_\nu$, $\widetilde{\nu} = (\widetilde{\nu}_1, \widetilde{\nu}_2, \ldots, \widetilde{\nu}_p)^T = (f_1(x_{\overline{\nu}}), f_2(x_{\overline{\nu}}), \ldots, f_p(x_{\overline{\nu}}))^T$, $F = g(\widetilde{\nu})$. Use rules (45) and (46) to produce H^{k+1} and B_{k+1}^δ and use formulas (49) and (50) to obtain set $\overline{B}_{k+1}^\delta$; let $S_{k+1}^\delta = B_{k+1}^\delta \setminus \overline{B}_{k+1}^\delta$, $\Xi^k = S_{k+1}^\delta \setminus T$. If $\Xi^k \neq \varnothing$, set $k = k+1$ and go to Step 1; otherwise, the algorithm terminates; let

Algorithm 2: Improved algorithm.

Using $\varepsilon \in (0,1), \delta = (1+\varepsilon)^{(1/\rho)}$ in Algorithm 1 and $(\varepsilon/2) < \ln(1+\varepsilon) < \varepsilon$, then there will be

$$\log_\delta \frac{u_i}{l_i} = \rho \log_{(1+\varepsilon)} \frac{u_i}{l_i} = \rho \frac{\ln(u_i/l_i)}{\ln(1+\varepsilon)} < \frac{2\rho \ln(u_i/l_i)}{\varepsilon}. \tag{61}$$

Then, by using the above formulas (55), (60), and (61), the upper limit of the number (expressed in $|B^\delta|$) of interior points of B^δ is

$$\left|B^\delta\right| \leq \left(\frac{2\widetilde{\alpha}\rho^2}{\varepsilon}\left[\ln 2\widetilde{q} + (n+1)\ln(n\lambda) - \ln\omega\right] + 1\right)^p, \tag{62}$$

in the utilized formula (55), (60), (61). From the above formula (62), we can see that the running time of Algorithm 1 is at most

$$O\left(\left(\frac{2\widetilde{\alpha}\rho^2}{\varepsilon}\left[(n+1)\ln(n\lambda) - \ln\omega\right] + 1\right)^p \cdot T(m+p,n)\right), \tag{63}$$

when the global ε–approximation solution is obtained, and then the proof of the conclusion is completed. □

Remark 2. Propositions 1 and 2 show that we can accelerate Algorithm 1 by removing the vertices of the small rectangle that needs not be considered, which leads to Algorithm 2 that is more resource-efficient than Algorithm 1; in other words, Algorithm 2 is an improvement on Algorithm 1. Then, the upper bound of the CPU running time required by Algorithm 2 is the same as that of Algorithm 1 in the most extreme cases (where acceleration techniques always fail). Therefore, Algorithm 2 is likewise a polynomial time approximation algorithm.

5. Numerical Experiments

This section will test the performance of the algorithm through several test problems. All of our testing procedures were performed via MATLAB (2012a) on computers with Intel(R) Core(TM)i5-2320, 3.00 GHz power processor, 4.00 GB memory, and Microsoft Win7 operating system.

Problem 1 (see [17, 18])

$$\begin{aligned}
&\min \begin{array}{l}(0.813396x_1 + 0.67440x_2 + 0.305038x_3 + 0.129742x_4 + 0.217796)\\ \times(0.224508x_1 + 0.063458x_2 + 0.932230x_3 + 0.528736x_4 + 0.091947)\end{array}\\
&\text{s.t.}\begin{cases}
0.488509x_1 + 0.063565x_2 + 0.945686x_3 + 0.210704x_4 \leq 3.562809,\\
-0.324014x_1 - 0.501754x_2 - 0.719204x_3 + 0.099562x_4 \leq -0.052215,\\
0.445225x_1 - 0.346896x_2 + 0.637939x_3 - 0.257623x_4 \leq 0.427920,\\
-0.202821x_1 + 0.647361x_2 + 0.920135x_3 - 0.983091x_4 \leq 0.840950,\\
-0.886420x_1 - 0.802444x_2 - 0.305441x_3 - 0.180123x_4 \leq -1.353686,\\
-0.515399x_1 - 0.424820x_2 + 0.897498x_3 + 0.187268x_4 \leq 2.137251,\\
-0.591515x_1 + 0.060581x_2 - 0.427365x_3 + 0.579388x_4 \leq -0.290987,\\
0.423524x_1 + 0.940496x_2 - 0.437944x_3 - 0.742941x_4 \leq 0.373620,\\
x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0.
\end{cases}
\end{aligned} \tag{64}$$

Problem 2 (see [17, 18])

$$\begin{aligned} &\min\ (3x_1 - 2x_2 - 2)^{(2/3)}(x_1 + 2x_2 + 2)^{(2/5)} \\ &\text{s.t.}\ \begin{cases} 2x_1 - x_2 \geq 2, x_1 - 2x_2 \leq 2, \\ x_1 + x_2 \leq 5, 3 \leq x_1 \leq 5, 1 \leq x_2 \leq 3. \end{cases} \end{aligned} \tag{65}$$

Problem 3 (see [8, 17, 18])

$$\begin{aligned} &\min\ (x_1 + x_2 + 1)^{2.5}(2x_1 + x_2 + 1)^{1.1}(x_1 + 2x_2 + 1)^{1.9} \\ &\text{s.t.}\ \begin{cases} x_1 + 2x_2 \leq 6, \\ 2x_1 + x_2 \leq 8, \\ 1 \leq x_1 \leq 3, \\ 1 \leq x_2 \leq 3. \end{cases} \end{aligned} \tag{66}$$

Problem 4 (see [20])

$$\begin{aligned} &\min\ (x_1 + x_2)(x_1 - x_2 + 7) \\ &\text{s.t.}\ \begin{cases} 2x_1 + x_2 \leq 14, \\ x_1 + x_2 \leq 10, \\ -4x_1 + x_2 \leq 0, \\ 2x_1 + x_2 \geq 6, \\ x_1 + 2x_2 \geq 6, \\ x_1 - x_2 \leq 3, \\ 1.99 \leq x_1 \leq 2.01, \\ 7.99 \leq x_2 \leq 8.01. \end{cases} \end{aligned} \tag{67}$$

Problem 5 (see [19])

$$\begin{aligned} &\min\ \left(c_1^T x + d_1\right)\left(c_2^T x + d_2\right) \\ &\text{s.t.}\ Ax = b, x \geq 0, \end{aligned} \tag{68}$$

where

$$A = \begin{pmatrix} 9 & 9 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 8 & 1 & 8 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 8 & 8 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & 1 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 7 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & 7 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{69}$$

$$\begin{aligned} b &= (81, 72, 72, 9, 9, 9, 8, 8)^T, \\ c_1 &= \left(1, 0, \frac{1}{9}, 0, 0, 0, 0, 0, 0, 0, 0\right)^T, \\ c_2 &= \left(0, 1, \frac{1}{9}, 0, 0, 0, 0, 0, 0, 0, 0\right)^T, \\ d_1 &= 0, \\ d_2 &= 0. \end{aligned}$$

Obviously, Problem 5 can be transformed into the following forms:

$$\begin{aligned} &\min\ \left(x_1 + \frac{1}{9}x_3\right)\left(x_2 + \frac{1}{9}x_3\right) \\ &\text{s.t.}\ \begin{cases} 9x_1 + 9x_2 + 2x_3 \leq 81, \\ 8x_1 + x_2 + 8x_3 \leq 72, \\ x_1 + 8x_2 + 8x_3 \leq 72, \\ 7x_1 + x_2 + x_3 \geq 9, \\ x_1 + 7x_2 + x_3 \geq 9, \\ x_1 + x_2 + 7x_3 \geq 9, \\ 0 \leq x_1 \leq 8, \\ 0 \leq x_2 \leq 8, \\ 0 \leq x_3 \leq 9. \end{cases} \end{aligned} \tag{70}$$

Problem 6 (see [8])

$$\begin{aligned} &\min\ (3x_1 - 4x_2 + 5)(x_1 + 2x_2 - 1)^{0.5}(2x_1 - x_2 + 4) \\ &\quad \times (x_1 - 2x_2 + 8)^{0.5}(2x_1 + x_2 - 1) \\ &\text{s.t.}\ \begin{cases} 5x_1 - 8x_2 \geq -24, \\ 5x_1 + 8x_2 \leq 44, \\ 6x_1 - 3x_2 \leq 15, \\ 4x_1 + 5x_2 \geq 10, \\ 1 \leq x_1 \leq 3, \\ 0 \leq x_2 \leq 1. \end{cases} \end{aligned} \tag{71}$$

Problem 7

$$\min \prod_{i=1}^{p} \left(c_i^T x_j + d_i\right)^{\alpha_i} \text{s.t.}\quad Ax \leq b, x \geq 0, \tag{72}$$

where $p \geq 2$, $c_i \in \mathbb{R}^n (i = 1, 2, \ldots, p)$ are pseudo-random numbers in [0, 1], $\alpha_i (i = 1, 2, \ldots, p)$ are pseudo-random numbers in [0.00001, 1], $d_i = 1$, constraint matrix elements a_{ij} are generated in [−1, 1] via $a_{ij} = 2 * \varpi - 1$, in which ϖ are pseudo-random numbers in [0, 1], and the right-hand side values are generated via $b_i = \sum_{j=1}^{n} a_{ij} + 2\beta_i$, in which β_i are pseudo-random numbers in [0, 1].

The numerical results in Tables 1 and 2 show that Algorithms 1 and 2 can effectively solve the three test problems known in the literature and get an approximate solution, so both algorithms are feasible.

Further, we do the corresponding random numerical experiments through Problem 7, which is utilized to explore the performance of the two algorithms. We determine the convergence accuracy of the algorithm to 0.05. For each set

Table 1: Comparison of results in Problems 1–6.

Problem	Reference	Optimal solution	Optimal optimum
1	Locatelli [17]	(1.3148, 0.1396, 0.0000, 0.4233)	0.890190
	Shen and Wang [18]	(1.3148, 0.1396, 0.0000, 0.4233)	0.890190
	Liu and Zhao [8]	(1.3148, 0.13955, 2.6891×10^{-14}, 0.42329)	0.890190
	Algorithms 1/2	(1.3148, 0.1396, 0.0000, 0.4233)	0.890190
2	Locatelli [17]	(3.000, 2.000)	5.014514
	Shen and Wang [18]	(3.000, 2.000)	5.009309
	Liu and Zhao [8]	(3.000, 2.000)	5.009309
	Algorithm 1	(3.000, 2.000)	5.009309
	Algorithm 2	(3.000, 2.000)	5.009309
3	Liu and Zhao [8]	(1, 1)	997.661265
	Locatelli [17]	(1, 1)	997.661265
	Shen and Wang [18]	(1, 1)	997.661265
	Algorithm 1/2	(1, 1)	997.661265
4	Shen and Hang [20]	(2, 8)	10
	Algorithm 1/2	(2, 8)	10
5	Zhang et al. [19]	(0.0, 8.0, 1.0, ...)	0.91235
	Algorithm 1	(0.0, 8.0, 1.0, ...)	0.91235
	Algorithm 2	(0.0, 8.0, 1.0, ...)	0.91235
6	Liu and Zhao [8]	(1.25, 1)	263.785989
	Algorithm 1	(1.25, 1)	263.785989
	Algorithm 2	(1.25, 1)	263.785989

Table 2: Comparison of results in Problems 1–6.

Problem	Reference	Iter	Time	ε
1	Locatelli [17]	404	9.606	0.05
	Shen and Wang [18]	3	0.047	0.05
	Algorithm 1/2	1	0.0149	0.05
2	Locatelli [17]	69	2.4960	0.15009
	Shen and Wang [18]	4	0.0800	0.15009
	Algorithm 1	6	0.1024	0.15009
	Algorithm 2	4	0.0657	0.15009
3	Locatelli [17]	5	1.126	0.2
	Shen and Wang [18]	4	0.085	0.2
	Algorithm 1/2	1	0.0116	0.2
4	Algorithm 1/2	1	0.0241	0.01
5	Algorithm 1	797	47.5367	0.2
	Algorithm 2	507	30.2111	0.2
6	Algorithm 1	63	59.4304	0.2
	Algorithm 2	37	35.6072	0.2

of fixed parameters (p, m, n), we run the two algorithms 10 times for numerical comparison, and the numerical results are given in Table 3. In Table 3, Avg (Std) time and Avg (Std) Iter represent the average (standard deviation) of the CPU running time and the average (standard deviation) of iterations, respectively, after the algorithm has run 10 times. Table 3 shows that the computation effect of Algorithm 2 is better than that of Algorithm 1, mainly because our acceleration technique plays a significant role by deleting the vertices of small rectangles that do not need to be considered. Hence, we believe that this acceleration technique may be generalized on other approximation algorithms such as [17, 18, 20].

Moreover, under the condition that the fixed parameters (p, m) are invariant, the CPU running time of the two algorithms will increase with the scale n of Problem 7. Under the condition that the prefixed parameters (m, n) are invariant, the CPU running time and iterations of the two

Table 3: Comparison of numerical results by using Problem 7.

(p, m, n)	Algorithm 1		Algorithm 2	
	Avg (Std) time	Avg (Std) Iter	Avg (Std) time	Avg (Std) Iter
(2, 10, 20)	2.9068 (2.8062)	75.8 (84.7700)	1.9686 (1.9352)	22.5 (27.3395)
(2, 20, 20)	2.3784 (3.1017)	52.6 (76.8936)	1.7129 (2.1472)	23.4 (35.1801)
(2, 22, 20)	0.8663 (0.9232)	18.1 (25.0257)	0.6568 (0.6239)	8 (10.0199)
(2, 20, 30)	6.2414 (6.3274)	165.2 (164.0334)	3.4923 (3.7124)	49 (61.764)
(2, 35, 50)	3.9868 (4.4041)	66.4 (78.2102)	3.3046 (3.9017)	32.3 (38.6886)
(2, 45, 60)	5.8908 (5.4016)	129.1 (125.2481)	3.7409 (3.3526)	40.5 (38.1084)
(2, 45, 100)	6.6579 (5.9685)	125.3 (123.7061)	4.2665 (3.6485)	40.2 (40.1343)
(2, 60, 100)	7.8626 (6.3057)	96.6 (99.4818)	4.5517 (2.8324)	26 (19.8343)
(2, 70, 100)	9.1245 (8.1057)	96.3 (104.6633)	5.0942 (3.3528)	23.6 (18.9430)
(2, 70, 120)	11.2742 (13.2311)	148 (215.2185)	6.0341 (5.5968)	35 (37.3256)
(2, 100, 10)	0.1877 (0.1300)	2.4 (2.9732)	0.1542 (0.0663)	1.3 (0.6403)
(2, 100, 50)	0.9029 (0.5392)	8.9 (7.0632)	0.6542 (0.3654)	3.9 (2.7730)
(2, 100, 100)	9.8811 (8.0793)	68.6 (55.5287)	6.9462 (6.3403)	24.1 (26.6813)
(2, 100, 150)	15.4331 (10.2573)	97.4 (75.1720)	9.8838 (6.2545)	30.8 (22.1260)
(2, 100, 200)	27.1157 (25.3267)	124.4 (130.8076)	18.9561 (16.8612)	49.2 (47.3810)
(2, 100, 250)	64.8144 (72.0125)	285.1 (353.7955)	40.3711 (41.0487)	91 (103.9576)
(2, 100, 300)	87.5572 (100.4846)	331 (398.8197)	55.5067 (64.5147)	117.2 (153.2434)
(2, 100, 400)	132.4251 (176.2381)	363.9 (581.9823)	87.0321 (97.4482)	130.7 (169.6585)
(2, 100, 500)	158.4767 (183.7060)	338.3 (493.9785)	111.0958 (106.7086)	133.8 (145.3470)
(2, 100, 700)	331.3275 (351.8534)	414.2 (546.8741)	272.7311 (264.9189)	227.1 (257.8927)
(2, 100, 1000)	1020.9318 (880.7910)	1063.6 (1019.7921)	778.8913 (638.1782)	522 (460.2479)
(3, 100, 10)	4.4724 (7.4341)	59.7 (117.2502)	3.6522 (5.7934)	35.2 (55.7867)
(3, 100, 50)	90.8139 (74.9843)	1062.4 (982.8277)	57.8342 (55.5978)	473.3 (564.6533)
(4, 100, 10)	75.4301 (189.1250)	1509.3 (4122.7180)	52.9122 (122.0553)	868.1 (2203.3283)

algorithms will grow with the number (p) of linear functions in the objective function of Problem 7.

6. Concluding Remarks

In this paper, we mainly propose two polynomial time approximation algorithms that can be utilized to solve the problem GLMP globally, where Algorithm 2 is obtained by accelerating Algorithm 1 by the proposed acceleration technique. The numerical results show that both algorithms are effective and feasible, but the overall calculation effect of Algorithm 2 is better than that of Algorithm 1, which shows that our acceleration technique is efficient and may be extended to some approximation algorithms such as [17, 18, 20].

Data Availability

All data and models generated or used during the study are described in Section 5 of this article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (grant no. 11961001), the Construction Project of First-Class Subjects in Ningxia Higher Education (NXYLXK2017B09), and the Major Proprietary Funded Project of North Minzu University (ZDZX201901).

References

[1] T. Matsui, "NP-Hardness of linear multiplicative programming and related problems," *Journal of Global Optimization*, vol. 9, no. 2, pp. 113–119, 1996.

[2] C. Maranas, I. Androulakis, C. Floudas et al., "Solving long-term financial planning problems via global optimization," *Journal of Economic Dynamics and Control*, vol. 21, no. 8-9, pp. 1405–1425, 1997.

[3] H. Konno, T. Kuno, and Y. Yajima, "Global minimization of a generalized convex multiplicative function," *Journal of Global Optimization*, vol. 4, no. 1, pp. 47–62, 1994.

[4] J. M. Mulvey, R. J. Vanderbei, and S. A. Zenios, "Robust optimization of large-scale systems," *Operations Research*, vol. 43, no. 2, pp. 264–281, 1995.

[5] R. Nicholas, P. Layard, and A. Walters, "Microeconomic theory," *Economica*, vol. 47, no. 186, p. 211, 1980.

[6] H. P. Benson, "Vector maximization with two objective functions," *Journal of Optimization Theory and Applications*, vol. 28, no. 2, pp. 253–257, 1979.

[7] D. Dennis, "Analyzing public inputs to multiple objective decisions on national forests using conjoint analysis," *Forest Ence*, vol. 44, no. 3, pp. 421–429, 1998.

[8] S. Liu and Y. Zhao, "An efficient algorithm for globally solving generalized linear multiplicative programming," *Journal of Computational and Applied Mathematics*, vol. 296, pp. 840–847, 2016.

[9] X. J. Liu, T. Umegaki, and Y. Yamamoto, "Heuristic methods for linear multiplicative programming," *Journal of Global Optimization*, vol. 15, no. 4, pp. 433–447, 1999.

[10] H. P. Benson and G. M. Boger, "Multiplicative programming problems: analysis and efficient point search heuristic," *Journal of Optimization Theory and Applications*, vol. 94, no. 2, pp. 487–510, 1997.

[11] P. Shen, X. Bai, and W. Li, "A new accelerating method for globally solving a class of nonconvex programming problems," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 71, no. 7-8, pp. 2866–2876, 2009.

[12] Y. Chen and H. Jiao, "A nonisolated optimal solution of general linear multiplicative programming problems," *Computers & Operations Research*, vol. 36, no. 9, pp. 2573–2579, 2009.

[13] C.-F. Wang, Y.-Q. Bai, and P.-P. Shen, "A practicable branch-and-bound algorithm for globally solving linear multiplicative programming," *Optimization*, vol. 66, no. 3, pp. 397–405, 2017.

[14] Y. Gao, C. Xu, and Y. Yang, "An outcome-space finite algorithm for solving linear multiplicative programming," *Applied Mathematics and Computation*, vol. 179, no. 2, pp. 494–505, 2006.

[15] H. Konno and T. Kuno, "Linear multiplicative programming," *Mathematical Programming*, vol. 56, no. 1-3, pp. 51–64, 1992.

[16] D. Depetrini and M. Locatelli, "A FPTAS for a class of linear multiplicative problems," *Computational Optimization and Applications*, vol. 44, no. 2, pp. 275–288, 2007.

[17] M. Locatelli, "Approximation algorithm for a class of global optimization problems," *Journal of Global Optimization*, vol. 55, no. 1, pp. 13–25, 2013.

[18] P. P. Shen and L. F. Wang, "A fully polynomial time approximation algorithm for generalized linear multiplicative programming," *Mathematica Applicata*, vol. 31, no. 1, pp. 208–213, 2018.

[19] B. Zhang, Y. Gao, X. Liu, and X. Huang, "Output-space branch-and-bound reduction algorithm for a class of linear multiplicative programs," *Mathematics*, vol. 8, no. 3, p. 315, 2020.

[20] P. Shen, B. Huang, and L. Wang, "Range division and linearization algorithm for a class of linear ratios optimization problems," *Journal of Computational and Applied Mathematics*, vol. 350, pp. 324–342, 2019.

[21] C.-F. Wang and S.-Y. Liu, "A new linearization method for generalized linear multiplicative programming," *Computers & Operations Research*, vol. 38, no. 7, pp. 1008–1013, 2011.

[22] P. P. Shen and X. K. Zhao, "A polynomial time approximation algorithm for linear fractionl programs," *Mathematica Applicata*, vol. 26, no. 2, pp. 355–359, 2011.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6692294, 7 pages
https://doi.org/10.1155/2021/6692294

Research Article

The Magnetic Bead Computing Model of the 0-1 Integer Programming Problem Based on DNA Cycle Hybridization

Rujie Xu,[1] Zhixiang Yin,[2] Zhen Tang,[1] Jing Yang,[1] Jianzhong Cui,[1,3] and Xiyuan Wang[1]

[1]*School of Mathematics and Big Data, AnHui University of Science & Technology, Huainan 232001, Anhui, China*
[2]*School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China*
[3]*School of Electronic and Information Engineering, AnHui University of Science & Technology, Huainan 232001, Anhui, China*

Correspondence should be addressed to Zhixiang Yin; zxyin66@163.com

Received 13 October 2020; Revised 4 November 2020; Accepted 4 February 2021; Published 16 February 2021

Academic Editor: Fazal Mahomed

Magnetic beads and magnetic Raman technology substrates have good magnetic response ability and surface-enhanced Raman technology (SERS) activity. Therefore, magnetic beads exhibit high sensitivity in SERS detection. In this paper, DNA cycle hybridization and magnetic bead models are combined to solve 0-1 integer programming problems. First, the model maps the variables to DNA strands with hairpin structures and weights them by the number of hairpin DNA strands. This result can be displayed by the specific binding of streptavidin and biotin. Second, the constraint condition of the 0-1 integer programming problem can be accomplished by detecting the signal intensity of the biological barcode to find the optimal solution. Finally, this model can be used to solve the general 0-1 integer programming problem and has more extensive applications than the previous DNA computing model.

1. Introduction

With the development of science and technology, traditional computing has been unable to meet people's requirements when dealing with massive data and information processing, and people have started to explore new fields of computing. Since Adleman proposed the use of DNA computing to solve the directed Hamilton path in 1994, DNA computing has received increasing attention from researchers [1]. In 2000, Head et al. proposed a new method of computing by using DNA plasmids and reported the NP-complete problem concerning the cardinality of the largest independent subset of the vertex set of the computing graph [2]. In 2011, Zhang et al. designed a DNA word set based on minimum free energy [3]. In 2017, Yin and Cui reported the integer programming problem based on the plasmid DNA computing model [4]. In 2018, Ramanamurthy introduced the basic structure of DNA and DNA processing tools [5]. In 2019, Tang established a dynamic NAND computing model using DNA origami [6]. In the same year, Yang et al. used DNA origami and hybridization chain reaction to solve a new computational model for solving the knapsack problem [7].

DNA cycle hybridization chain reaction is a process of alternating hybridization of two DNA molecules with different hairpin structures induced under the induction of a trigger strand. This process is spontaneous and does not require the involvement of enzymes. With the development of science and technology, DNA cycle hybridization chain reaction has been applied to many fields, such as biosensing, biomedicine, proteins, and others. In 2004, Dirks first proposed the concept and indicated that DNA can be used as an amplified transducer for biosensing applications [8]. In 2016, Guo proposed a new chemical immunoassay method for signal amplification that can detect multiple tumor biomarkers simultaneously [9]. In the same year, Yang designed an Aptamer-Binding Directed DNA Origami Pattern for logic gates [10]. In 2018, Li et al. proposed a method for label-free lighting of fluorescent sensors using hybrid chain reaction and DNA triple-strand assembly [11]. In 2018, Xiao designed multiple chemiluminescence imaging and used it for sensitive screening and detection of protein biomarkers through the use of DNA microarray and hybridization chain reaction amplification integration induced by adjacent binding [12].

0-1 integer programming is a special type of integer programming problem, and its variable values are only 0 and 1. It is widely used in a variety of problems, such as line design, backpack problems, and dispatch problems [13–17]. In 2006, Yin et al. designed a molecular beacon model for solving integer linear programming problems [13]. In 2010, Huang used the advantage of DNA tiles to build a molecular computing system to solve the 0-1 programming problem [14]. In 2017, Li devised a new DNA computing method to solve the 0-1 programming problem. The method is based on a self-assembled nanoparticle probe, which reduces the cost of the model and improves the sensitivity and accuracy of detection [17].

In this paper, based on DNA cycle hybridization chain reaction, a magnetic bead model is constructed to solve the general 0-1 integer programming problem. This magnetic bead model maps variables into the hairpin structure DNA strands and maps weights according to the number of hairpin structure DNA strands. This structure can be displayed according to the specific combination of streptavidin and biotin, and SERS detection has been used [18–20]. The general structure of this paper is as follows. First, the basic principle of the DNA cycle hybridization chain reaction and the general form of the 0-1 integer programming problem are given. Second, the magnetic bead model is constructed, and the algorithm steps of the model and a concrete example are given. Finally, this paper uses Visual DSD software to simulate and analyze the optimal solution of the 0-1 integer programming problem and then provides the conclusion.

2. DNA Cycle Hybridization and the 0-1 Integer Programming Problem

2.1. DNA Cycle Hybridization. DNA cycle hybridization refers to the process in which DNA molecules with complementary base sequences form hydrogen bonds between base pairs to achieve a stable structure. DNA cycle hybridization induces two different types of hairpin structure DNA by using a trigger strand, making the two hairpin structures open alternately and then forming a double-stranded DNA product with a gap. The process of obtaining the DNA product does not require the involvement of enzymes. The reaction principle of DNA cycle hybridization is shown in Figure 1. Two different types of hairpin structures, H_1 and H_2, coexist stably in solution without any reaction. H_1 consists of four parts: $5' - l - m - n - m^* - 3'$ (m and m^* base pairs are complementary). It is called the stem of the hairpin. Region n is called the loop of the hairpin structure H_1, and l is the single-stranded sticky end extended from the stem of H_1. Similarly, H_2 consists of four parts: $5' - n^* - m - l^* - m^* - 3'$ (m and m^* base pairs are complementary). It is called the stem of the hairpin, and n^* is the single-stranded sticky end extended from the stem of H_2. At the same time, $l - l^*$, $m - m^*$, and $n - n^*$ satisfy the principle of complementary base pairing. The trigger strand T is a single strand of DNA composed of two parts ($5' - m^* - l^* - 3'$). When the trigger strand T is added to the solution, the trigger strand T will have base complementary pairing with the sticky end of $l - m$ of the hairpin structure of H_1. The resulting hairpin structure H_1 is opened, exposing the gap $n - m^*$. This exposed gap region $n - m^*$ happens to have complementary base pairing with $n^* - m$ from the H_2 region. Opening the structure of H_2 to expose the area $l^* - m^*$, the exposed area $l^* - m^*$ from H_2 will continue to pair with the next base complement of H_1. Opening the structure of H_1 to expose the area $n - m^*$, this exposed gap region will continue to have complementary base pairing with $n^* - m$ from the H_2 region. Opening the structure of H_2, H_1 and H_2 are turned on alternately in turn until the sum in the solution is used up. A double-stranded DNA product that hybridizes H_1 and alternately H_2 with a gap is generated.

2.2. Integer Programming Problem. 0-1 integer programming is a special type of integer programming problem, and its variable values are only 0 and 1. The general form of 0-1 programming is given as follows:

$$\max(\min) z = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$
$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \le (=, \ge) b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \le (=, \ge) b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \le (=, \ge) b_m \\ x_1, x_2, \ldots, x_n = 0 \text{ or } 1 \\ a_{ij} \text{ and } c_i \text{ are any integer} \\ i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n. \end{cases} \tag{1}$$

The most commonly used method for solving the 0-1 integer programming problem is the exhaustive method (also known as the forced search method), which traverses the entire search space, but this method is time- and labor-consuming.

In this paper, the general 0-1 integer programming problem is solved based on the DNA cyclc hybridization chain reaction, which is a generalization of the assignment problem.

The algorithm steps of the general 0-1 integer programming problem are discussed as follows:

Step 1. All possible solutions to the problem with variables equal to 0 or 1 are given

Step 2. According to the Raman signal intensity, the nonfeasible solutions in the first constraint condition are eliminated, and the feasible solutions are retained

Step 3. Repeat step 2 with the reserved feasible solutions; then, all nonfeasible solutions can be eliminated, and all feasible solutions of the problem can be obtained

Step 4. The corresponding objective function values of each feasible solution are compared to obtain the optimal solution of the objective function.

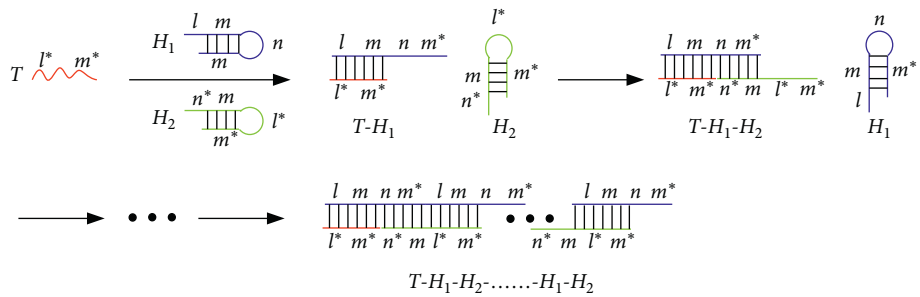


Figure 1: The basic principle of DNA cycle hybridization.

3. Magnetic Bead Model of the 0-1 Integer Programming Problem Based on DNA Cycle Hybridization

3.1. Building a Magnetic Bead Computing Model. In this paper, we study a new magnetic bead computing model for 0-1 integer programming problems. The magnetic bead computing model effectively utilizes the specific binding effect of streptavidin and biotin through DNA cycle hybridization technology and Raman technology to detect the signal released by the biological barcode. Table 1 shows the coding sequences of the three DNA molecules. Figure 2 shows the DNA cycle hybridization process.

As shown in Figure 2, at the optimal experimental conditions of 37°C and a pH of 7.4, the concentration of the captured stranded DNA was 1.0×10^{-7}M [21]. The single strand of capture DNA was fixed on the magnetic bead by means of the amide bond between -COOH modified by the magnetic bead and -NH2 on the DNA strand. The single strand of capture DNA fixed on the magnetic bead was complementarily paired with the base at the sticky end of the strand of the hairpin DNA1 strand, thus opening the hairpin structure DNA1 strand. When the hairpin structure DNA1 is opened, it is complementary with the sticky end base of the hairpin structure DNA2, thus opening the hairpin structure of DNA2. After the hairpin structure DNA2 is opened, it continues to be complementary to the sticky end of the hairpin structure of DNA1. In this way, the hairpin structure DNA1 strand and DNA2 strand cycle hybridizes successively, forming "magnetic bead-capture DNA-DNA1-DNA2-DNA1-...-DNA2-DNA1-DNA2," a special double-stranded DNA molecule. Until DNA1 and DNA2 are consumed in the solution, the sticky end of DNA1 and the sticky end of DNA2 are both modified by biotin and bind to the strepavidin-modified nanobiotic barcode specifically to achieve signal release.

In summary, for the 0-1 integer programming problem with n variables $(x_1, x_2, \ldots, x_n)$ and m constraint equations, the specific algorithm of the general 0-1 integer programming computing model is as follows:

Step 1. First, for n variables in each constraint condition, n magnetic beads with capture DNA were designed (magnetic beads with different radii represent different variables). Second, two types of hairpin structure DNA strands were designed, which were named DNA1 and DNA2. The capture DNA fixed on the magnetic bead can open the hairpin structure DNA1. A gap appeared after the hairpin structure DNA1 was opened, which could be further opened to design the hairpin structure DNA2. In this way, DNA1, DNA2, DNA1, DNA2... cyclically cross each other in turn. Until DNA1 and DNA2 in the solution are consumed (the specific coding sequence design of DNA1 and DNA2 is shown in Table 1), when $x_i = 1$, the sticky ends of the DNA1 strand and the sticky ends of the DNA2 strand were modified with biotin; when $x_i = 0$, the sticky ends of the DNA1 strand and the sticky ends of the DNA2 strand did not need to be modified with biotin, as shown in Figure 3.

Step 2. First, a proper number of biological barcodes were placed in the data pool and mixed evenly. Second, a set of test tubes was prepared for each constraint condition, and each set of test tubes had 2^n test tubes (where n represents the number of variables in the constraint condition). Finally, equal amounts of the solution were placed in the desired tubes.

Step 3. For the first constraint, according to the number of possible solutions k, take out the test tubes according to step 2 and group them and place a magnetic bead with capture DNA in each test tube of each group. After that, put equal amounts of DNA1 and DNA2 into the solution according to the weight coefficients of the variables in the constraint condition. That is, the total amounts of DNA1 and DNA2 are the same as the weight coefficients of the variables. At the same time, according to the characteristics of DNA cycle hybridization, we alternately put DNA1 and DNA2 into the solution every time and put the DNA1 strand first.

Step 4. When the biological barcode in the solution is combined with the sticky ends of the biotin of DNA1 and DNA2, the cycle hybridization signal will be amplified, and the feasibility solution will be judged by the intensity of amplification of the cycle hybridization signal. Here, it is stated that, when there is no biotin at the sticky ends of DNA1 and DNA2, the signal intensity is 0, when 1 biological barcode in the solution binds to 1 biotin, the signal intensity is 1, and so on, and when biological barcodes bind to biotin in the solution, the signal intensity is a, where a represents the coefficient in front of each variable, namely, the weight.

Table 1: The coding sequences of DNA molecules.

Name	Coding sequence
Capture DNA	3-ATAAGGGGGAAAAGATTTGATTTGTT-NH_2-5
DNA1	5-Biotin- TATTCCCCCTTTTCTAAACTAAACAA GCTATTGTTTAGTTTAGAAAAGGG-3
DNA2	3-Biotin-CGATCCCTTTTCTAAACTAAACAAATAAGGTTGTTTAGTTTAGAAAAGGG-5

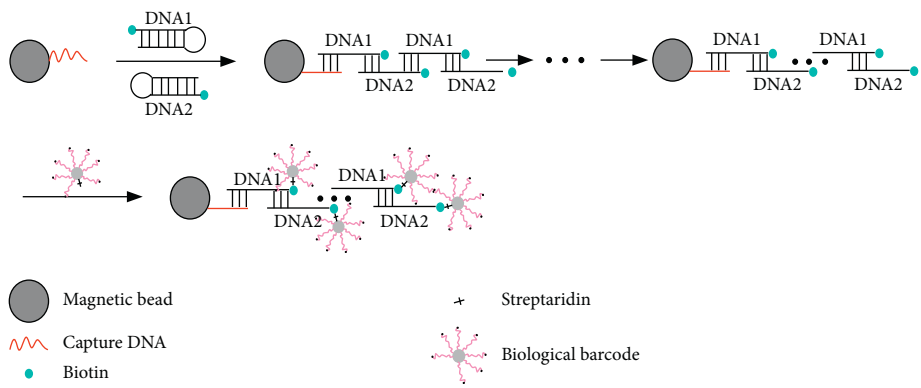


Figure 2: The process of DNA cycle hybridization.

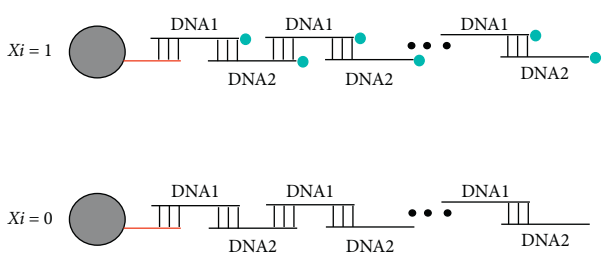


Figure 3: The structure diagram of variables $x_i = 1$ and.$x_i = 0$.

Step 5. Find a feasible solution satisfying the first constraint condition by detecting the number of biological barcodes.

Step 6. Repeat steps 4–6 above for the feasible solution obtained from the previous constraint condition, and we can obtain the feasible solution that satisfies all of the constraints.

Step 7. Calculate each feasible solution corresponding to the objective function value and, finally, judge the optimal integer programming solution.

3.2. Example Analysis. A general 0-1 integer programming problem is discussed in detail as follows:

$$\min w = 4x_1 + 3x_2 + 5x_3$$
$$\begin{cases} 3x_1 + 2x_2 + 4x_3 \geq 5 \\ 2x_1 + 3x_2 \leq 3 \\ x_2 + 2x_3 \leq 2 \\ x_1, x_2, x_3 = 0 \text{ or } 1. \end{cases} \quad (2)$$

Step 1. For the variables in each constraint condition, magnetic beads with capture DNA were designed. In this problem, 3, 2, and 2 magnetic beads with capture DNA were designed for the three constraints, with magnetic bead radii of 2 nm, 4 nm, and 6 nm, respectively. Then, two types of hairpin structure DNA strands were designed, known as DNA1 and DNA2. When the value of variable x_i is 1, the sticky ends of the DNA1 and DNA2 strands are modified with biotin. When the value of the variable x_i is 0, the sticky ends of the DNA1 and DNA2 strands do not need to be modified with biotin.

Step 2. An appropriate amount of the biological barcode was placed in the solution and mixed evenly. Three sets of test tubes were prepared, and the number of each set of test tubes was 8, 4, and 4. The correct amount and equal amount of solution were placed into the test tubes.

Step 3. All possible solutions of the objective function variables are denoted as $1(0,0,0)$, $2(0,0,1)$, $3(0,1,0)$, $4(0,1,1)$, $5(1,0,0)$, $6(1,0,1)$, $7(1,1,0)$, and $8(1,1,1)$. For the first constraint condition, prepare 8 sets of test tubes, which are labeled 1, 2, 3, 4, 5, 6, 7, and 8, corresponding to the 8 possible solutions of the previous step. There are three test tubes in each set of test tubes, each of which is put into a magnetic bead with captured DNA, which are recorded as x_1, x_2, x_3, respectively, and the radii of the magnetic beads are 2 nm, 4 nm, and 6 nm, Put DNA1 and DNA2 into the respective test tubes according to the x_1, x_2, x_3 coefficients in the constraint condition.

Step 4. The specific process is shown in Figure 4.

Step 5. The signal intensities in the 8 test tubes are 0, 4, 2, 6, 3, 7, 5, and 9. The feasible solutions that satisfy the first constraint condition are $4(0,1,1)$, $6(1,0,1)$, $7(1,1,0)$, and $8(1,1,1)$.

Step 6. Because the second constraint does not involve x_3, we only need to consider x_1 and x_2. For the feasible solutions obtained in step 6, the 4th, 6th, 7th, and 8th groups

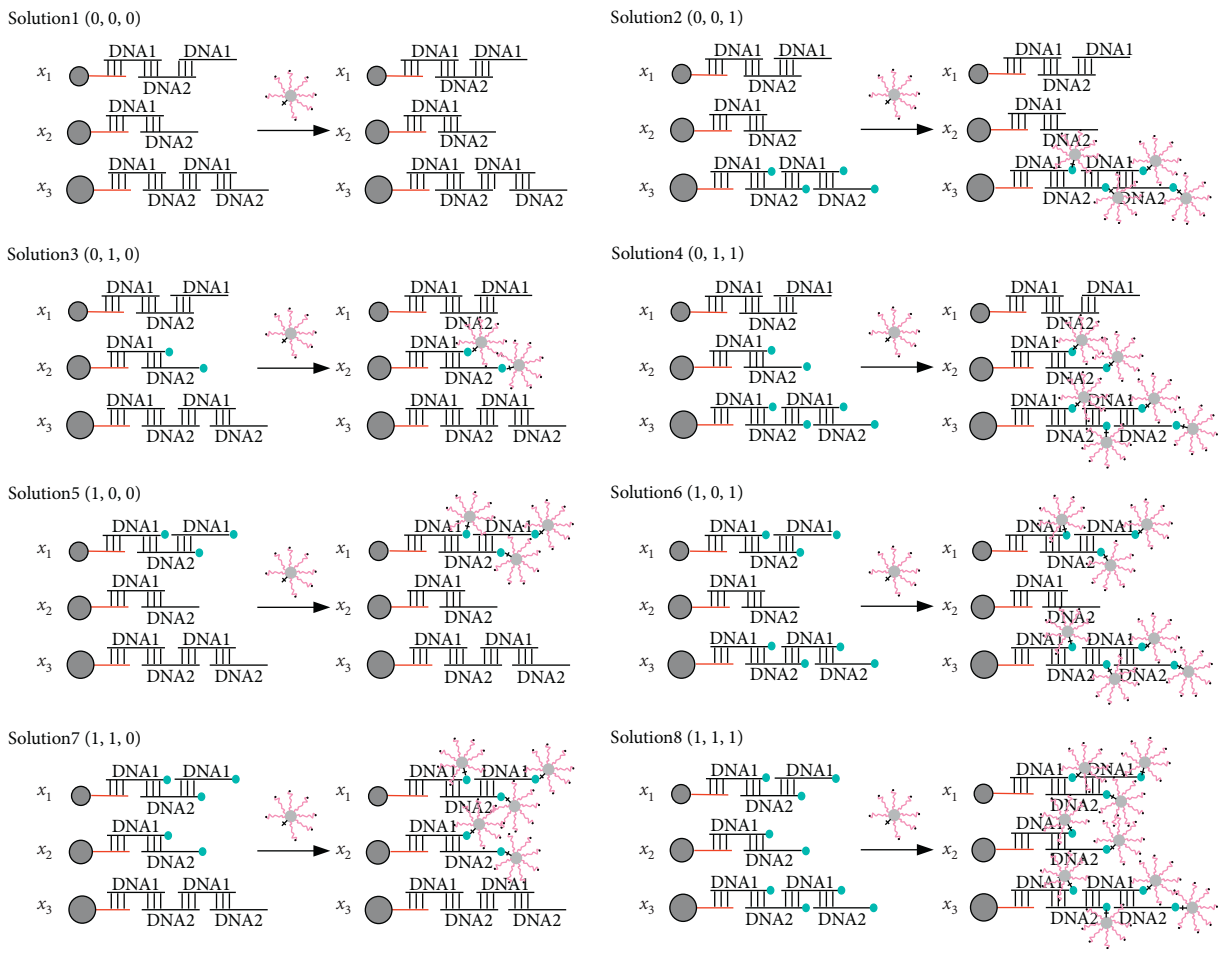


Figure 4: The structure diagram of 8 solutions.

of solutions, the values are 4 (0, 1), 6 (1, 0), 7 (1, 1), and 8 (1, 1). Among them, the 7th and 8th solutions have the same values, and only the 7th solution (1,1) is considered here. Continue to steps 4 and 5, as shown in Figure 5.

The signal intensities of these three groups of test tubes, 4, 6, and 7, are 3, 2, and 5, respectively. Only group 4 (0,1) and 6 (1,0) test tubes meet the second constraint. Thus, the feasible solutions satisfying the first two constraints are 4 (0,1,1) and 6 (1,0,1).

Because the third constraint condition does not involve variable x_1, we only need to consider the values x_2 and x_3, and the values are 4 (1, 1) and 6 (0, 1). Continue to steps 4 and 5. The specific process is shown in Figure 6 below.

Step 7. Finally, group 6 of solutions (1, 0, 1) is a feasible solution that satisfies all constraints. Substituting the feasible solution into the objective function, the minimum objective function of the 0-1 integer programming problem can be obtained as 9.

4. Discussion

Visual DSD is a simulation software commonly used in DNA computing and hybridization chain reaction. This paper uses Visual DSD software to simulate and analyze the optimal solution of the 0-1 integer programming problem. The optimal solution of the example integer programming problem is $(x_1, x_2, x_3) = (1, 0, 1)$. For variables $x_1 = 1$, add the hairpin structure DNA1 strand and DNA2 strand, and because the reaction is just started, the concentration of reactants is higher and the reaction speed is faster. The concentration of the hairpin structure DNA1 and DNA2 strands decreases rapidly in a short time and eventually gradually approaches 0. For sp5, the intermediate product of the reaction, because the cycle hybridization reaction is carried out step by step, the concentration of the strand first increases and then decreases before finally approaching 0. The concentration of the final product sp4 gradually increases and finally tends to be stable. The specific reaction process is shown in Figure 7. The simulation results show that the model is feasible and consistent with the expected results.

Previous models, such as the DNA origami base, circular logic gate, and others, cannot solve the weighted integer programming problem, which increases the understanding space virtually. The magnetic bead model proposed in this paper, which can solve the 0-1 integer programming problem with weight, can solve the general 0-1 integer programming problem, so it is more widely used.

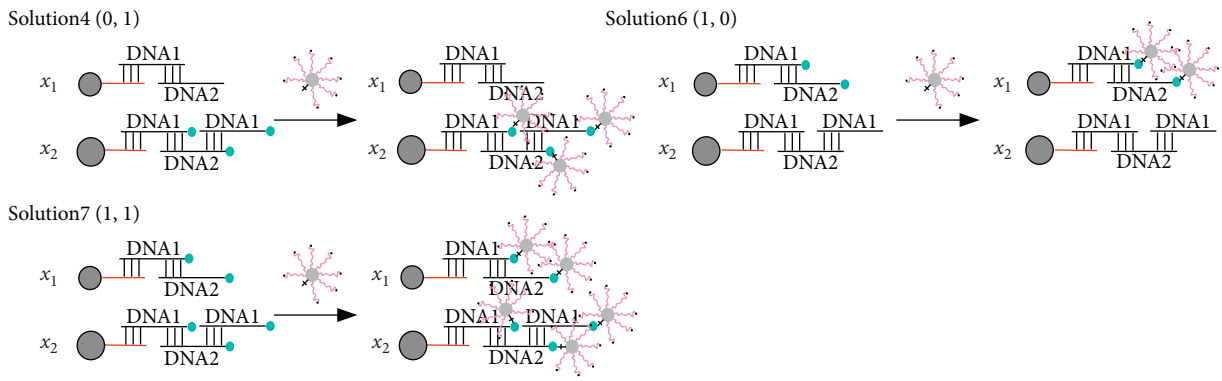


Figure 5: Diagram of the solution under the second constraint.

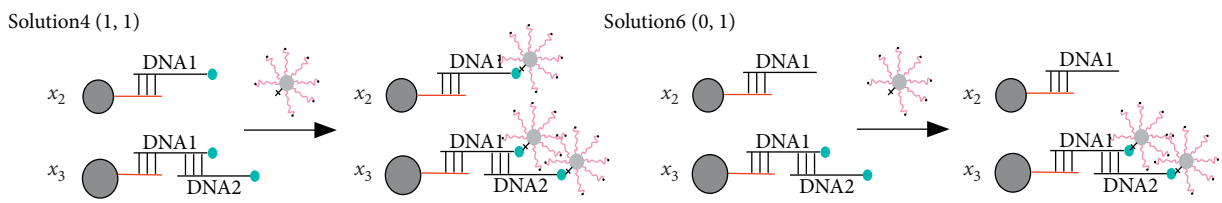


Figure 6: Diagram of the solution under the third constraint.

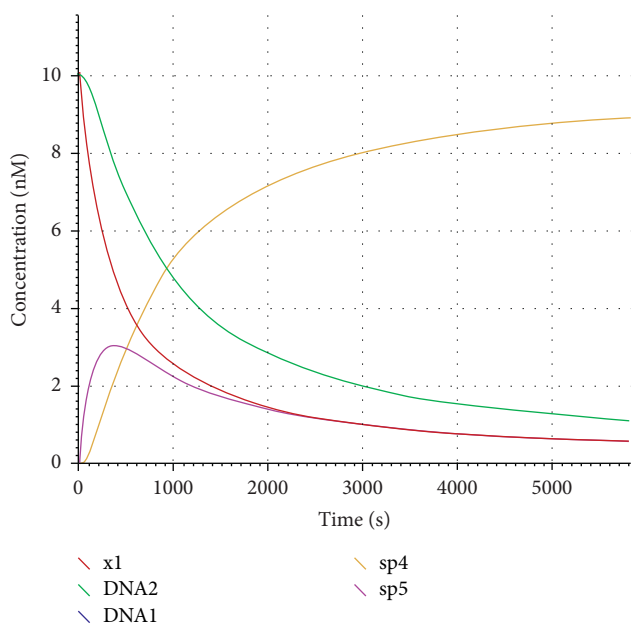


Figure 7: Diagram of model simulation.

5. Conclusion

In this paper, a magnetic bead model for solving the 0-1 integer programming problem was established based on the DNA cycle hybridization chain reaction and the specific binding effect of streptavidin and biotin. Compared with the previous DNA computing model, this model has the following advantages. First, there is no requirement for enzymes in the operation process, which can reduce the experimental cost and improve the versatility of the model. Second, the intensity of the signal is used to judge the feasibility of the solution. This can improve the accuracy and practicability of the detection results. Finally, this model can be used to solve the general 0-1 integer programming problem and has more extensive applications than the previous DNA computing model. However, this method still has some shortcomings, such as a large number of steps and long operation time. Therefore, these aspects still need to be studied further.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The project was supported by the National Natural Science Foundation of China (nos. 62072296, 61672001, 61702008, and 11971302), Natural Science Foundation of Anhui Province (no. 1808085 MF193), Foreign Visit and Study Project of Outstanding Young Talents in Colleges and Universities (no. gxgwfx2019015), Sub-Project of CST Forward Innovation Project (no. 18163ZT00500901), and Anhui Postdoctoral Fund (no. 2019B331).

References

[1] L. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, no. 5187, pp. 1021–1024, 1994.

[2] T. Head, G. Rozenberg, R. S. Bladergroen, C. K. D. Breek, P. H. M. Lommerse, and H. P. Spaink, "Computing with DNA

by operating on plasmids," *Biosystems*, vol. 57, no. 2, pp. 87–93, 2000.

[3] Q. Zhang, B. Wang, X. Wei et al., "DNA word set design based on minimum free energy," *IEEE Transactions on Nanobioence*, vol. 9, no. 4, pp. 273–277, 2011.

[4] Z. Yin, J. Cui, and J. Yang, "Integer programming problem based on plasmid DNA computing model," *Chinese Journal of Electronics*, vol. 26, no. 6, pp. 1284–1288, 2017.

[5] S. V. Ramanamurthy, K. Hyndhavi, and B. Sruthi Sai Nirmala, "DNA computing-the future of computing," *Journal of Innovation in Computer Sceince Engineering*, vol. 8, no. 1, pp. 18–22, 2018.

[6] Z. Tang, Z.-X. Yin, X. Sun, J.-Z. Cui, J. Yang, and R.-S. Wang, "Dynamically NAND gate system on DNA origami template," *Computers in Biology and Medicine*, vol. 109, pp. 112–120, 2019.

[7] J. Yang, Z. Yin, Z. Tang, K. Huang, J. Cui, and X. Yang, "Search computing model for the knapsack problem based on DNA origami," *Materials Express*, vol. 9, no. 6, pp. 553–562, 2019.

[8] R. M. Dirks and N. A. Pierce, "From the Cover: triggered amplification by hybridization chain reaction," *Proceedings of the National Academy of Sciences*, vol. 101, no. 43, pp. 15275–15278, 2004.

[9] J. Guo, J. Wang, J. Zhao, Z. Guo, and Y. Zhang, "Ultrasensitive multiplexed immunoassay for tumor biomarkers based on DNA hybridization chain reaction amplifying signal," *ACS Applied Materials & Interfaces*, vol. 8, no. 11, pp. 6898–6904, 2016.

[10] J. Yang, S. Jiang, X. Liu, L. Pan, and C. Zhang, "Aptamer-binding directed DNA origami pattern for logic gates," *ACS Applied Materials & Interfaces*, vol. 8, no. 49, pp. 34054–34060, 2016.

[11] Z. Li, L. Tingting, S. Ruidi et al., "A label-free light-up fluorescent sensing platform based upon hybridization chain reaction amplification and DNA triplex assembly," *Talanta*, vol. 189, pp. 137–142, 2018.

[12] Q. Xiao, J. Wu, P. Dang, and H. Ju, "Multiplexed chemiluminescence imaging assay of protein biomarkers using DNA microarray with proximity binding-induced hybridization chain reaction amplification," *Analytica Chimica Acta*, vol. 1032, pp. 130–137, 2018.

[13] Z. X. Yin, J. Z. Cui, J. Yang, and X. Yang, *BDNA Computing Model of the Integer Linear Programming Problem Based on Molecular Beacon*, Springer, Berlin, Germany, 2006.

[14] X. Zheng, J. Yang, C. Zhou, C. Zhang, Q. Zhang, and X. Wei, "Allosteric DNAzyme-based DNA logic circuit: operations and dynamic analysis," *Nucleic Acids Research*, vol. 47, no. 3, pp. 1097–1109, 2019.

[15] Y. Huang, Z. Cheng, J. Xu, X. Shi, and K. Zhou, "Solving 0-1 planning problem based on self-assembly of DNA tiles," *Journal of Computational and Theoretical Nanoscience*, vol. 7, no. 5, pp. 826–830, 2010.

[16] J. Yang, C. Zhang, S. Liu, H. Xia, and J. Xu, "A molecular computing model for 0-1 programming problem using DNA nanoparticles," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 10, pp. 2380–2384, 2013.

[17] F. Li, "DNA computation based on self-assembled nanoparticle probes for 0-1 integer programming problem," *Science*, vol. 268, no. 5210, pp. 542–546, 2017.

[18] A. M. Michaels, M. Nirmal, and L. E. Brus, "Surface enhanced Raman spectroscopy of individual rhodamine 6G molecules on large Ag nanocrystals," *Journal of the American Chemical Society*, vol. 121, no. 43, pp. 9932–9939, 1999.

[19] H. Zhang, C. Fang, and S. Zhang, "An autonomous bio-barcode DNA machine for exponential DNA amplification and its application to the electrochemical determination of adenosine triphosphate," *Chemistry-A European Journal*, vol. 16, no. 41, pp. 12434–12439, 2010.

[20] B. Wang, Q. Zhang, and X. Wei, "Tabu variable neighborhood search for designing DNA barcodes," *IEEE Transactions on NanoBioscience*, vol. 19, no. 1, pp. 127–131, 2020.

[21] Y. Lu, G. L. Liu, and L. P. Lee, "High-density silver nanoparticle film with temperature-controllable interparticle spacing for a tunable surface enhanced Raman scattering substrate," *Nano Letters*, vol. 5, no. 1, pp. 5–9, 2005.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6643349, 13 pages
https://doi.org/10.1155/2021/6643349

Research Article

Extinction Moment for a Branching Tree Evolution with Birth Rate and Death Rate Both Depending on Age

Xi Hu,[1] Yun-Zhi Yan,[2] Zhong-Tuan Zheng,[1] Hong-Yan Li,[3] and Hong-Yan Zhao[1]

[1]*School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China*
[2]*School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance, Shanghai 201620, China*
[3]*School of Management Studies, Shanghai University of Engineering Science, Shanghai 201620, China*

Correspondence should be addressed to Xi Hu; xih_xih@163.com

Received 12 November 2020; Revised 26 December 2020; Accepted 28 January 2021; Published 11 February 2021

Academic Editor: Jiyuan Tao

In this paper, a branching tree evolution is established, in which the birth rate and the death rate are both dependent on node's age. The extinction probability and the t-pre-extinction (extinct before time t) probability are studied, by which the distribution of the extinction moment can be given. The analytical formula and the approximation algorithm for the distribution of extinction moment are given; furthermore, the analytical formula and the approximation algorithm of extinction probability are given, and a necessary and sufficient condition of extinction with probability 1 is given. It is the first time to study the distribution of extinction time for the branching process with birth rate and the death rate both depending on node's age, and the results will do great help in the theory of branching process. It is expected to be applied in the fields of biology, genetics, medicine, epidemiology, demography, nuclear physics, actuarial mathematics, algorithm, and data structures, etc.

1. Introduction

The classical biological reproduction model G-W branching process [1] has been extended to different biological reproduction models, such as branching processes in random environments [2–4] and branching population evolution models [5–11]. The age-dependent branching process was introduced by Bellman and Harris [6]. In branching models, the population extinction problem is one of the primary research contents. Many problems in branching models related to the population extinction are studied, but the distribution of extinction moment is hardly involved. In this paper, a branching tree evolution is established, in which the birth rate and death rate are both dependent on node's age. The extinction probability and the t-pre-extinction (extinct before time t) probability are studied, by which the distribution of the extinction moment can be given.

The paper is organized as follows. The model is described and the existence theorem is presented in Section 2. In Section 3, the extinction probability is studied, and the analytical formula and the approximation algorithm of the extinction probability are given. A necessary and sufficient condition of extinction with probability 1 is also given. In Section 4, the t-pre-extinction probability is studied, the iterative integral equation with unique solution is established, which is satisfied by the t-pre-extinction probability, and the analytical formula and the approximation algorithm of t-pre-extinction probability are given. The stochastic order of extinction moment is studied in Section 5. The conclusions are presented in Section 6.

2. Description and Existence Theorem for the Model

In this paper, based on the mechanism of asexual reproduction of biological population, a continuous time random graph evolution is constructed, in which a node's birth rate and death rate are both dependent on the node's age.

Given a population is composed of biological individuals (nodes). The evolution of the population is based on the following basic assumptions:

(1) All nodes in the population are homogeneous and mutually independent

(2) The node's death rate in the population is a non-negative function $\alpha(\cdot)$ dependent on the node's age, such that $\int_0^\infty \alpha(t)\mathrm{d}t = +\infty$

(3) The node's birth rate in the population is a non-negative function $\beta(\cdot)$ dependent on the node's age

(4) Conditioned under a node being alive, the node's reproduction behaviors in the future are conditional independent

(5) Conditioned under a node being alive, the node's death is conditional independent with the node's reproduction

(6) At initial time $t = 0$, there is only one initial node in the population (this condition is not essential, only for convenience of presentation)

(7) In addition to the initial node, each of other nodes in the population has only one parent node

Based on the above assumptions, the branching tree evolution is described as follows.

Given a node i in the population, its age is s at time t. For a sufficiently small period $\Delta t > 0$, conditioned under node i being alive at time t, the conditional probability for node i being dead in the period $[t, t+\Delta t)$ is $\alpha(s)\Delta t + o(\Delta t)$, the conditional probability for node i producing one child node in the period $[t, t+\Delta t)$ is $\beta(s)\Delta t + o(\Delta t)$, and the conditional probability for node i producing more than one child node in the period $[t, t+\Delta t)$ is $o(\Delta t)$.

In the population, if node j is a child of node i, then there is a directed link from node i to node j. When at least one of the parent and child dies, the link between them is a virtual (dotted) line, and the dead node is called a virtual node. Otherwise, it is called a real node, and so on. At time $t \geq 0$, all nodes (real and virtual) and directed links (real and virtual) construct a directed random tree, denoted by $G_t(\cdot)$. And thus, the process of reproduction is an evolution of random trees, denoted by $\{G_t(\cdot)\}_{t\geq 0}$. As the evolution is characterized by the birth rate $\beta(\cdot)$ and the death rate $\alpha(\cdot)$, therefore, the model is referred to as "branching tree evolution with birth rate and death rate both depending on age," denoted by $\{G_t(\beta(\cdot), \alpha(\cdot))\}_{t\geq 0}$.

According to the definition of the model, $\forall t \geq 0$, the number of offspring born in period $(0, t]$ is finite, and no more than one offspring will be born at the same time. Therefore, the initial node and all its offspring nodes can be ordered as $1, 2, \ldots, n, \ldots$ according to the order of birth time.

$\forall n \geq 1$, denote

$$\begin{aligned}
&\vec{n} = (1, 2, \ldots, n),\\
&\vec{f}(n) = (f_1, f_2, \ldots, f_n): \quad f_1 = 0, 1 \leq f_k \leq k-1, 2 \leq k \leq n,\\
&\vec{i}(n) = (i_1, i_2, \ldots, i_n): \quad i_k \in \{0, 1\}, 1 \leq k \leq n,\\
&\vec{b}(n) = (b_1, b_2, \ldots, b_n): \quad b_i \in R_+, 1 \leq i \leq n,\\
&\qquad \text{and } 0 = b_1 < b_2 < \cdots < b_n,
\end{aligned} \tag{1}$$

where $\vec{n}$ is the vector of the labeled nodes; $\vec{f}(n)$ is the vector of the adjacency relation (parent-child relation) between nodes: $f_1 = 0$ means that the initial node has no parent node. $f_k = j, 1 \leq j \leq k-1$ indicates that node j is the parent of node $k, 2 \leq k \leq n$; $\vec{i}(n)$ is the vector of node's alive-death status: $i_k = 1$ denotes that node k is alive and $i_k = 0$ denotes that node k is dead, $1 \leq k \leq n$; and $\vec{b}(n)$ is the birth time vector: $b_1 = 0$ represents there is an initial node at time $t = 0$, b_j is the time when node j is born, $2 \leq j \leq n$, and $b_1 < b_2 < \cdots < b_n$ implies that no more than one node is born at the same time.

$\forall n \geq 1$, denote

$$\begin{aligned}
\overline{F}_n &= \left\{\vec{f}(n)\right\},\\
\overline{E}_n &= \left\{\vec{i}(n)\right\},\\
\overline{B}_n &= \left\{\vec{b}(n)\right\},\\
C_{3\times n} &= (\vec{n}, \vec{f}(n), \vec{i}(n))^T,\\
S_n^{(3)} &= \{\vec{n}\} \times \overline{F}_n \times \overline{E}_n,\\
C_{4\times n} &= (\vec{n}, \vec{f}(n), \vec{i}(n), \vec{b}(n))^T,\\
S_n &= \{\vec{n}\} \times \overline{F}_n \times \overline{E}_n \times \overline{B}_n,\\
S &= \bigcup_{n=1}^{\infty} S_n.
\end{aligned} \tag{2}$$

$\forall C_{4\times n} = (\vec{n}, \vec{f}(n), \vec{i}(n), \vec{b}(n))^T$, denote $I(C_{4\times n}) = \{k: i_k = 1\}$. $\forall k \in I(C_{4\times n})$, denote

$$\begin{aligned}
D_k(C_{4\times n}) = \Big\{&C_{4\times(n+1)} = (\overrightarrow{n+1}, \vec{f}(n+1), \vec{r}(n+1), \vec{b}(n+1))^T:\\
&\vec{b}(n+1) = \left(\vec{b}(n), b_{n+1}\right); \vec{f}(n+1) = \left(\vec{f}(n), f_{n+1}\right), f_{n+1} = k; r_j \leq i_j, 1 \leq j \leq n\Big\}.
\end{aligned} \tag{3}$$

$\forall C_{4\times n} = (\overrightarrow{n}, \overrightarrow{f}(n), \overrightarrow{i}(n), \overrightarrow{b}(n))$, and $b_n \le s < t$, define the function $f_n^{(k)}(s,t,C_{4\times(n+1)}|C_{4\times n})$ on S_{n+1}. $\forall C_{4\times(n+1)} = (\overrightarrow{n+1}, \overrightarrow{f}(n+1), \overrightarrow{r}(n+1), \overrightarrow{b}(n+1))$,

$$\begin{aligned} f_n^{(k)}\left(s,t,C_{4\times(n+1)}|C_{4\times n}\right) &= \widetilde{I}_{D_k(C_{4\times n})}\left(C_{4\times(n+1)}\right)\left(\prod_{\substack{j\in I(C_{4\times n})\\ j\neq k, r_j=0}} \int_s^t e^{-\int_{s-b_j}^{y-b_j}\beta(u)\mathrm{d}u}\alpha(y)e^{-\int_s^y \alpha(u)\mathrm{d}u}\mathrm{d}y\right) \\ &\times\left(\prod_{\substack{j\in I(C_{4\times n})\\ j\neq k, r_j=1}} e^{-\int_{s-b_j}^{t-b_j}\beta(u)\mathrm{d}u}e^{-\int_s^t \alpha(u)\mathrm{d}u}\right) \\ &\times \beta(b_{n+1}-b_k)g_1(t,b_k,r_k)g_2(t,b_{n+1},r_{n+1}), \end{aligned} \tag{4}$$

where $\widetilde{I}_A(\cdot)$ is a indicative function, and

$$\begin{aligned} g_1(t,b_k,r_k) &= \begin{cases} e^{-\int_s^t \alpha(u)\mathrm{d}u}e^{-\int_{s-b_j}^{t-b_j}\beta(u)\mathrm{d}u}, & r_k = 1, \\ e^{-\int_{s-b_k}^{b_{n+1}-b_k}\beta(u)\mathrm{d}u}\int_{b_{n+1}}^t e^{-\int_{b_{n+1}-b_k}^{y-b_k}\beta(u)\mathrm{d}u}\alpha(y)e^{-\int_s^y \alpha(u)\mathrm{d}u}\mathrm{d}y, & r_k = 0, \end{cases} \\ g_2(t,b_k,r_{n+1}) &= \begin{cases} e^{-\int_{b_{n+1}}^t \alpha(u)\mathrm{d}u}e^{-\int_{b_{n+1}}^t \beta(u)\mathrm{d}u}, & r_{n+1} = 1, \\ \int_{b_{n+1}}^t e^{-\int_{b_{n+1}}^y \beta(u)\mathrm{d}u}\alpha(y)e^{-\int_{b_{n+1}}^y \alpha(u)\mathrm{d}u}\mathrm{d}y, & r_{n+1} = 0. \end{cases} \end{aligned} \tag{5}$$

$\forall C_{4\times n} = (\overrightarrow{n}, \overrightarrow{f}(n), \overrightarrow{i}(n), \overrightarrow{b}(n))$, and $b_n \le s < t$, define the function $f_n(s,t,C_{4\times(n+1)}|C_{4\times n})$ on S_{n+1}:

$$f_n\left(s,t,C_{4\times(n+1)}|C_{4\times n}\right) = \sum_{k\in I(C_{4\times n})} f_n^{(k)}\left(s,t,C_{4\times(n+1)}|C_{4\times n}\right). \tag{6}$$

Let $\mathrm{d}b_{n+1}$ be a Lebesgue measure on $(s,t]$, for a given $\overrightarrow{b}(n)$, and $b_n \le s$, then $\delta_{\left\{\overrightarrow{b}(n)\right\}} \times \mathrm{d}b_{n+1}$ is a measure on $(\overline{B}_{n+1}, \mathscr{B}(\overline{B}_{n+1}))$. Let $\mu_{n+1}(\cdot)$ be a count measure on $(S_{n+1}^{(3)}, \mathscr{B}(S_{n+1}^{(3)}))$, denote $\nu_{n+1}(\cdot) = \mu_{n+1} \times (\delta_{\left\{\overrightarrow{b}(n)\right\}} \times \mathrm{d}b_{n+1})(\cdot)$, and then $\nu_{n+1}(\cdot)$ is a measure on $(S_{n+1}, \mathscr{B}(S_{n+1}))$. Define $\forall D_{n+1} \in \mathscr{B}(S_{n+1})$,

$$Q_n\left(s,t,D_{n+1}|C_{4\times n}\right) = \int_{D_{n+1}} f_n\left(s,t,C_{4\times(n+1)}|C_{4\times n}\right)\nu_{n+1}\left(\mathrm{d}C_{4\times(n+1)}\right). \tag{7}$$

Then, $\forall 0 \le s < t$, $C_{4\times n} \in S_n$, $Q_n(s,t,\cdot|C_{4\times n})$ is a measure on $(S_{n+1}, \mathscr{B}(S_{n+1}))$, $\forall 0 \le s < t$, $D_{n+1} \in \mathscr{B}(S_{n+1})$, $Q_n(s,t,D_{n+1}|\cdot)$ is a measurable function on $(S_n, \mathscr{B}(S_n))$.

Let

$$\begin{aligned} A_{2\times n} &= \begin{pmatrix} \overrightarrow{n} \\ \overrightarrow{f}(n) \end{pmatrix}, \\ B_{2\times n} &= \begin{pmatrix} \overrightarrow{i}(n) \\ \overrightarrow{b}(n) \end{pmatrix}, \\ C_{4\times n} &= \begin{pmatrix} A_{2\times n} \\ B_{2\times n} \end{pmatrix}. \end{aligned} \tag{8}$$

Let $N(t)$ be the number of nodes in the random branching tree $G_t(\beta(\cdot),\alpha(\cdot))$, then $G_t(\beta(\cdot),\alpha(\cdot))$ can be expressed by a $2\times N(t)$ matrix, i.e.,

$$G_t(\beta(\cdot),\alpha(\cdot)) = A_{2\times N(t)}. \tag{9}$$

The birth time and the alive-death status of the $N(t)$ nodes in $G_t(\beta(\cdot),\alpha(\cdot))$ can be expressed by the $2\times N(t)$ matrix $B_{2\times N(t)}$, and denote

$$X_t = \begin{pmatrix} A_{2\times N(t)} \\ B_{2\times N(t)} \end{pmatrix} = C_{4\times N(t)}. \tag{10}$$

We have the following theorem.

Theorem (existence) $\{G_t(\beta(\cdot),\alpha(\cdot))\}_{t\geq 0}$ is the marginal process of the nonhomogeneous Markov process $\{X_t\}_{t\geq 0}$ in the state space S, where the transfer function of $\{X_t\}_{t\geq 0}$ is $\forall D\in\mathscr{B}(S), 0\leq s<t, n\geq 1, C_{4\times n} = (\vec{n}, \vec{f}(n), \vec{i}(n), \vec{b}(n))^T \in S_n$, here $D_m \in \mathscr{B}(S_m)$

$$\begin{aligned} &P(X(t)\in D|X(s)=C_{4\times n}) \\ &= \sum_{m=1}^{\infty} P(X(t)\in D, N(t)=m|X(s)=C_{4\times n}) \\ &= \sum_{m=1}^{\infty} P(X(t)\in D_m|X(t)=C_{4\times n}) \\ &= P(X(t)\in D_n|X(s)=C_{4\times n}) + \sum_{m=n+1}^{\infty} P(X(t)\in D_m|X(s)=C_{4\times n}), \end{aligned} \tag{11}$$

where

$$P(X(t)\in D_n|X(s)=C_{4\times n}) = \sum_{\vec{r}(n)\in\widehat{D}_n} \prod_{\substack{j\in I(C_{4\times n}) \\ r_j=0}} \int_s^t e^{-[\lambda(y-b_j)-\lambda(s-b_j)]}\alpha e^{-\alpha(y-s)}\mathrm{d}y \times \prod_{\substack{j\in I(C_{4\times n}) \\ r_j=1}} e^{-[\lambda(t-b_j)-\lambda(s-b_j)]}e^{-\alpha(t-s)}, \tag{12}$$

$$\begin{aligned} &P(X(t)\in D_{n+k}|X(s)=C_{4\times n}) = P(X(t)\in D_{n+k}|X(s)=C_{4\times n}) \\ &= \int_{S_{n+1}}\cdots\int_{S_{n+k-1}}\int_{D_{n+k}} Q_n(s, b_{n+1}, \mathrm{d}C_{4\times(n+1)}|C_{4\times n})\cdots \\ &\quad\times Q_{n+k-2}(b_{n+k-2}, b_{n+k-1}, dC_{4\times(n+k-1)}|C_{4\times(n+k-2)}) \\ &\quad\times Q_{n+k-1}(b_{n+k-1}, t, \mathrm{d}C_{4\times(n+k)}|C_{4\times(n+k-1)}), \quad k\geq 1, \end{aligned} \tag{13}$$

where $\widehat{D}_n = \left\{\vec{r}(n)\colon (\vec{n}, \vec{f}(n), \vec{r}(n), \vec{b}(n)) \in D_n\right\}$. Substituting (12) and (13) into (11), the transfer function of $\{X_t\}_{t\geq 0}$ is obtained.

It is not difficult to prove that S is a Borel subset of the separable complete distance space R^{∞}, and the existence theorem of $\{X_t\}_{t\geq 0}$ can be proved by the existence theorem of Markov process. $\{G_t(\beta(\cdot),\alpha(\cdot))\}_{t\geq 0}$ is a marginal process of $\{X_t\}_{t\geq 0}$, and thus, the existence of $\{G_t(\beta(\cdot),\alpha(\cdot))\}_{t\geq 0}$ is proved.

3. The Extinction Probability

Define

$$\begin{aligned} T(\omega) &= \inf\{t>0\colon \text{population extinction in the period}[0,t)\}; \\ P(t) &= P\{T(\omega)\leq t\}, \quad 0\leq t<\infty; \\ P(\infty) &= P\{T(\omega)<\infty\}. \end{aligned} \tag{14}$$

$T(\omega)$ is called the extinction moment, at which the population extinct. The probability $P(t)$ is called t-pre-extinction probability, which is the probability of the population extinct before time t, and the probability $P(\infty)$ is called extinction probability. The distribution of extinction moment $T(\omega)$ is given by t-pre-extinction probability and $P(T(\omega)=\infty) = 1-P(T(\omega)<\infty) = 1-P(\infty)$. If $P(\infty)=1$, i.e., $P(T(\omega)=\infty)=0$. Then, $T(\omega)$ is a real-valued random variable, so the t-pre-extinction probability $P(t), t\geq 0$ is the distribution function of $T(\omega)$.

In this section, the extinction probability for the branching tree evolution $\{G_t(\beta(\cdot),\alpha(\cdot))\}_{t\geq 0}$ is studied, the analytical formula and the approximation algorithm of extinction probability are given, and a necessary and sufficient condition of extinction with probability 1 is also given.

Let $\eta(t)$ be the number of nodes that are alive in the population at time t, then $P(t) = P(\eta(t)=0), 0\leq t<\infty$. Obviously, $\{\eta(s)=0\}\subseteq\{\eta(t)=0\}$ when $s\leq t$, and $P(\infty) = \lim_{t\longrightarrow\infty} P(\eta(t)=0)$.

It is obvious that $P(t)$ has the following properties:

(1) $\forall 0 < s < t, 0 = P(0) < P(s) < P(t) < P(\infty)$

(2) $P(t)$ is continuous on $[0, \infty)$

Lemma 1. *Given a node i in the population, its lifespan is Y, and then Y has the probability density function:*

$$f_Y(t) = \begin{cases} \alpha(t)e^{-\int_0^t \alpha(u)\mathrm{d}u}, & t \geq 0, \\ 0, & t < 0, \end{cases} \tag{15}$$

where $\alpha(\cdot)$ is the node's death rate.

It is easy to prove Lemma 1.

Theorem 1. *The following iterative integral equation is satisfied by the t-pre-extinction probability $P(t), 0 < t < \infty$*

$$P(t) = \int_0^t e^{-\int_0^s (1 - P(t-u))\beta(u)\mathrm{d}u} \cdot \alpha(s)e^{-\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s, \quad 0 < t < \infty, \tag{16}$$

where $\beta(\cdot)$ is the node's birth rate and $\alpha(\cdot)$ is the node's death rate.

Proof. $\forall 0 < t < \infty, 0 < P(t) < 1$, Let Y be the node's lifespan, then we get

$$\begin{aligned} P(t) &= P(\eta(t) = 0) = P(\eta(t) = 0, Y \leq t) \\ &= \int_0^t P(\eta(t) = 0|Y = s) \cdot \alpha(s)e^{-\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s. \end{aligned} \tag{17}$$

In the following, we first calculate $P(\eta(t) = 0|Y = s)$. Equally divide the interval $[0, s]$ into n intervals, denote $\delta = (s/n)$, and let A_k be the random event: the initial node produces a child node in the period $(k\delta, (k+1)\delta]$ and the offspring of this child node extinct before time t or the initial node does not produce a child node in the period $(k\delta, (k+1)\delta]$, $0 \leq k \leq n-1$.

When δ is sufficiently small, the probability that the initial node does not produce a child node in the period $(k\delta, (k+1)\delta]$ is $1 - \beta(k\delta) \cdot \delta + o(\delta)$; the probability of producing more than one child node is $o(\delta)$; the probability of producing one child node is $\beta(k\delta) \cdot \delta + o(\delta)$, and this child node's offspring extinct before time t with probability $P(t - k\delta), 0 \leq k \leq n-1$, so

$$P(A_k|Y = s) = \beta(k\delta) \cdot \delta \cdot P(t - k\delta) + 1 - \beta(k\delta) \cdot \delta + o(\delta), \quad 0 \leq k \leq n-1. \tag{18}$$

Noting the independent assumptions of the model, we have

$$\begin{aligned} P(\eta(t) = 0|Y = s) &= \lim_{\delta \longrightarrow 0} P\left(\bigcap_{k=0}^{n-1} A_k|Y = s\right) = \lim_{\delta \longrightarrow 0} \prod_{k=0}^{n-1} P(A_k|Y = s) \\ &= \lim_{\delta \longrightarrow 0} \prod_{k=0}^{n-1} [\beta(k\delta) \cdot \delta \cdot P(t - k\delta) + (1 - \beta(k\delta) \cdot \delta) + o(\delta)] \\ &= \lim_{\delta \longrightarrow 0} \prod_{k=0}^{n-1} [1 - \beta(k\delta) \cdot (1 - P(t - k\delta)) \cdot \delta + o(\delta)]. \end{aligned} \tag{19}$$

Then,

$$\begin{aligned} \ln P(\eta(t) = 0|Y = s) &= \lim_{\delta \longrightarrow 0} \sum_{k=0}^{n-1} \ln[1 - \beta(k\delta) \cdot (1 - P(t - k\delta)) \cdot \delta + o(\delta)] \\ &= -\lim_{\delta \longrightarrow 0} \sum_{k=0}^{n-1} \beta(k\delta) \cdot (1 - P(t - k\delta)) \cdot \delta + o(\delta). \end{aligned} \tag{20}$$

That is

$$P(\eta(t) = 0|Y = s) = e^{-\int_0^s (1 - P(t-u))\beta(u)\mathrm{d}u}. \tag{21}$$

So

$$P(t) = \int_0^t e^{-\int_0^s (1 - P(t-u))\beta(u)\mathrm{d}u} \cdot \alpha(s)e^{-\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s. \tag{22}$$

Thus, the theorem is proved.

Denote

$$g(x) = \int_0^\infty e^{-\int_0^s (1-x)\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s, \quad 0 \le x \le 1. \tag{23}$$

□

Lemma 2. *The extinction probability $P(\infty)$ is a solution of the equation*

$$x = g(x), \quad 0 \le x \le 1. \tag{24}$$

Proof. $0 < t_0 < t < \infty$,

$$\begin{aligned}
&|P(t) - g(P(\infty))| \\
&= \left| \int_0^t e^{-\int_0^s (1-P(t-u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s - \int_0^\infty e^{-\int_0^s (1-P(\infty))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \right| \\
&\le \left| \int_0^{t_0} e^{-\int_0^s (1-P(t-u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s - \int_0^{t_0} e^{-\int_0^s (1-P(\infty))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \right| \\
&\quad + \left| \int_{t_0}^t e^{-\int_0^s (1-P(t-u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s - \int_{t_0}^t e^{-\int_0^s (1-P(\infty))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \right| \\
&\quad + \int_t^\infty e^{-\int_0^s (1-P(t-u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s,
\end{aligned} \tag{25}$$

$\forall \varepsilon > 0, \exists t_0, 0 < t_0 < t$, and let t be large enough, such that

$$\begin{aligned}
&\left| \int_0^{t_0} e^{-\int_0^s (1-P(t-u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s - \int_0^{t_0} e^{-\int_0^s (1-P(\infty))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \right| < \frac{\varepsilon}{3}, \\
&\left| \int_{t_0}^t e^{-\int_0^s (1-P(t-u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s - \int_{t_0}^t e^{-\int_0^s (1-P(\infty))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \right| < \frac{\varepsilon}{3}, \\
&\int_t^\infty e^{-\int_0^s (1-P(\infty))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s < \frac{\varepsilon}{3},
\end{aligned} \tag{26}$$

i.e., $\lim_{t \longrightarrow \infty} P(t) = g(P(\infty))$, which imply $P(\infty) = g(P(\infty))$.

Thus, Lemma 2 is proved.

It is easy to prove that the function $g(x)$ has the following properties. □

Lemma 3

(1) $g(0) > 0, g(1) = 1$

(2) $g(x)$ *is increasing on* $[0, 1]$

(3) $g(x)$ *is a strictly concave function on* $[0, 1]$

Theorem 2. *The extinction probability $P(\infty)$ is the smallest solution of the equation*

$$x = g(x), \quad 0 \le x \le 1. \tag{27}$$

Proof. By Lemma 3, $g(x)$ is a strictly concave function on $[0, 1]$, and thus, $\tilde{g}(x) = g(x) - x$ is also a strictly concave function on $[0, 1]$. It is easy to see that any strictly concave function has at most two different roots in its definition domain; hence, $\tilde{g}(x) = 0$ has at most two different solutions on $[0, 1]$, one of which is $x = 1$. Let $x = q$ be the smallest solution of the equation $x = g(x)$.

(1) If $q = 1$, since $q = 1$ is the smallest solution of the equation, then the equation has no solution in $(0, 1)$. But, by Lemma 2, $P(\infty)$ is the solution of the equation, infer that $P(\infty) = 1$, i.e., $P(\infty)$ is the smallest solution of the equation $x = g(x)$.

(2) $0<q<1$, let x such that $q<x<1$, since q is the unique solution of the equation $x=g(x)$ in $(0,1)$, it is easy to see that $x>g(x)$.

It is easy to prove $\forall t>0, P(t)\le q$. In fact, suppose contrarily $\exists t>0$, such that $P(t)>q$, then

$$P(t)>g(P(t)). \tag{28}$$

Noting that $P(t)$ is increasing, then

$$\begin{aligned} P(t) &= \int_0^t e^{-\int_0^s (1-P(t-u))\beta(u)\mathrm{d}u}\cdot\alpha(s)e^{-\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s \\ &\le \int_0^t e^{-\int_0^s (1-P(t))\beta(u)\mathrm{d}u}\cdot\alpha(s)e^{-\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s \\ &\le \int_0^\infty e^{-\int_0^s (1-P(t))\beta(u)\mathrm{d}u}\cdot\alpha(s)e^{-\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s \\ &= g(P(t)). \end{aligned} \tag{29}$$

This contradicts $P(t)>g(P(t))$, so the above assumption is not true. By Lemma 2, $P(\infty)$ is the solution of the equation, i.e., $P(\infty)=g(P(\infty))\le q$. By the fact that q is the smallest solution of the equation, imply that $P(\infty)=q$. So $P(\infty)$ is the smallest solution of the equation. Therefore, the theorem is proved. □

Corollary 1. *If $\beta(\cdot)=\beta>0$, then*

$$P(\infty)=\frac{\alpha}{\beta}\wedge 1. \tag{30}$$

Furthermore, $P(\infty)=1\Leftrightarrow\alpha\ge\beta$.

Proof. If $\beta(\cdot)=\beta>0$, then

$$\begin{aligned} g(x) &= \int_0^\infty e^{-(1-x)\int_0^s \beta(u)\mathrm{d}u}\cdot\alpha e^{-\alpha s}\mathrm{d}s \\ &= \int_0^\infty \alpha e^{-[(1-x)\beta+\alpha]s}\mathrm{d}s \\ &= \frac{\alpha}{(1-x)\beta+\alpha}. \end{aligned} \tag{31}$$

Let $g(x)=x$, that is,

$$\begin{aligned} \frac{\alpha}{(1-x)\beta+\alpha} &= x, \\ \beta x^2-(\beta+\alpha)x+\alpha &= 0. \end{aligned} \tag{32}$$

Obviously, $\Delta=(\beta+\alpha)^2-4\beta\alpha=(\beta-\alpha)^2\ge 0$; the two roots of the above equation are as follows:

$$x_{1,2}=\frac{(\beta+\alpha)\pm|\beta-\alpha|}{2\beta}. \tag{33}$$

Then,

$$P(\infty)=\frac{\alpha}{\beta}\wedge 1, \tag{34}$$

thus,

$$P(\infty)=1\Leftrightarrow\alpha\ge\beta. \tag{35}$$

The proof is completed.

As a consequence of Theorem 2, a sufficient condition for $P(\infty)=1$ is given. □

Corollary 2. *$\forall s>0$, if $\int_0^s\beta(u)du\le\int_0^s\alpha(u)du$, then $P(\infty)=1$.*

Proof. According to the assumptions, there is

$$\begin{aligned} g(x) &= \int_0^\infty e^{-(1-x)\int_0^s \beta(u)\mathrm{d}u}\cdot\alpha(s)e^{-\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s \\ &\ge \int_0^\infty e^{-(1-x)\int_0^s \alpha(u)\mathrm{d}u}\cdot\alpha(s)e^{-\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s \\ &= \frac{1}{2-x}\int_0^\infty (2-x)\alpha(s)e^{-(2-x)\int_0^s \alpha(u)\mathrm{d}u}\mathrm{d}s \\ &= \frac{1}{2-x}. \end{aligned} \tag{36}$$

If $0<x\le 1$ is a solution of the equation $x=g(x)$, then x satisfies

$$g(x)=x\ge\frac{1}{2-x}. \tag{37}$$

Thus, $x^2-2x+1\le 0$, i.e., $(x-1)^2\le 0$, so $x=1$, deduce $P(\infty)=1$.

Corollary 2 shows that when the death rate is greater than the birth rate, the population is certainly extinct, which is intuitive. □

Corollary 3. *Let $\{G_t(\beta_1(\cdot),\alpha(\cdot))\}_{t\ge 0}$ and $\{G_t(\beta_2(\cdot),\alpha(\cdot))\}_{t\ge 0}$ be two branching tree evolutions with different birth rates and the same death rate. The corresponding extinction probabilities are denoted by $P_1(\infty)$ and $P_2(\infty)$, respectively. If $\beta_1(u)\ge\beta_2(u), u\ge 0$, then $P_1(\infty)\le P_2(\infty)$.*

Proof. $\forall 0\le x\le 1$, $g_i(x)=\int_0^\infty e^{-(1-x)\int_0^s\beta_i(u)\mathrm{d}u}\cdot\alpha(s)e^{-\int_0^s\alpha(u)\mathrm{d}u}\mathrm{d}s, i=1,2$.

If $\beta_1(u)\ge\beta_2(u), u\ge 0$, then by the definition of $g_i(x), i=1,2$, it is easy to see that $g_1(x)\le g_2(x), 0\le x\le 1$, and $g_1(x)-x\le g_2(x)-x$, $0\le x\le 1$.

Since $g_i(x)-x$ is a continuous function with at least one smallest root on $[0,1](i=1,2)$, and $0<g_1(0)\le g_2(0)$, therefore, the smallest root of $g_1(x)-x$ on $(0,1]$ is less than or equal to the smallest root of $g_2(x)-x$ on $(0,1]$, and by Theorem 2, there is

$$P_1(\infty)\le P_2(\infty). \tag{38}$$

The proof is completed.

Corollary 3 implies that, for two models with the same death rate, the higher the birth rate is, the lower the extinction probability is, which is intuitive. □

Corollary 4. *Let* $\{G_t(\beta(\cdot),\alpha_1(\cdot))\}_{t\geq 0}$ *and* $\{G_t(\beta(\cdot)\alpha_2(\cdot))\}_{t\geq 0}$ *be two branching tree evolutions with different death rates but with the same birth rate. The corresponding extinction probabilities are denoted by* $P_1(\infty)$ *and* $P_2(\infty)$*, respectively,* $\forall s\geq 0$*, if* $\int_0^s \alpha_1(u)du \leq \int_0^s \alpha_2(u)du$*, then* $P_1(\infty)\leq P_2(\infty)$.

Proof. Denote $\overline{F}_i(s) = 1 - F_i(s) = e^{-\int_0^s \alpha_i(u)du}$, $i = 1,2$, $\forall s\geq 0$, if $\int_0^s \alpha_1(u)du \leq \int_0^s \alpha_2(u)du$, then $\overline{F}_1(s)\geq \overline{F}_2(s)$.

So $\int_0^\infty A(s)dF_1(s)\leq \int_0^\infty A(s)dF_2(s)$ for any decreasing function $A(\cdot)$.

The corresponding functions to $g(x)$ are denoted by $g_1(x)$ and $g_2(x)$, respectively. Noting that $e^{-(1-x)\int_0^s \beta(u)du}$ is a decreasing function with s, so by the definition of

$$g_i(x) = \int_0^\infty e^{-(1-x)\int_0^s \beta(u)du} dF_i(s), \quad i = 1,2. \tag{39}$$

imply $g_1(x)\leq g_2(x)$, $0\leq x\leq 1$, then $g_1(x) - x \leq g_2(x) - x$, $0\leq x\leq 1$,

Since $g_i(x) - x$ is a continuous function with at least one smallest root on $[0,1]$, $i = 1,2$, and $0 < g_1(0)\leq g_2(0)$, therefore, the smallest root of $g_1(x) - x$ on $(0,1]$ is less than or equal to the smallest root of $g_2(x) - x$ on $(0,1]$, and by Theorem 2,

$$P_1(\infty)\leq P_2(\infty). \tag{40}$$

Corollary 4 shows that, for two models with the same birth rate, the randomly longer the lifespan is, the smaller the extinction probability is, which is intuitive. □

Theorem 3. $P(\infty) = 1 \Leftrightarrow g'(x) < 1, 0 < x < 1$.

Proof

(1) Sufficiency: assume $g'(x) < 1, 0 < x < 1$, let $\widetilde{g}(x) = g(x) - x$, $0\leq x\leq 1$, then $\widetilde{g}'(x) = g'(x) - 1 < 0, 0 < x < 1$; i.e., $\widetilde{g}(x)$ is the decreasing function on $[0,1]$. Noting that $\widetilde{g}(0) = g(0) > 0, \widetilde{g}(1) = g(1) - 1 = 0$, thus $x = 1$ is the smallest root of $\widetilde{g}(x)$ on $[0,1]$, and by Theorem 2, we get $P(\infty) = 1$.

(2) Necessity: assume $P(\infty) = 1$, because $g(x)$ is a strictly concave function on $[0,1]$, $\widetilde{g}(x)$ is also a strictly concave function on $[0,1]$, and $\widetilde{g}(0) = g(0) > 0$. In addition, by Theorem 2 and the assumptions, it is obvious that $x = 1$ is the smallest root of $\widetilde{g}(x)$ on $[0,1]$, so $\widetilde{g}(x)$ is decreasing on $[0,1]$. Hence, $\widetilde{g}'(x) < 0$, that is, $g'(x) < 1, 0 < x < 1$.

Thus, the theorem is proved.

For $g(x) = \int_0^\infty e^{-(1-x)\int_0^s \beta(u)du}\cdot \alpha(s)e^{-\int_0^s \alpha(u)du}ds, 0\leq x \leq 1$ introduced above, noting that $\forall 0\leq x\leq 1$, $0 < g(x)\leq 1$, denote

$$\begin{aligned} g_0(x) &= x, \quad 0\leq x\leq 1,\\ g_1(x) &= g(x),\\ g_n(x) &= g(g_{n-1}(x)), \quad n\geq 2, \end{aligned} \tag{41}$$

i.e., $g_n(\cdot)$ is the n times iteration of $g(\cdot)$, and then, there is the following conclusion. □

Theorem 4. $\forall 0\leq x < 1$, *there is* $\lim_{n\longrightarrow\infty} g_n(x) = P(\infty)$.

Proof

(1) If $P(\infty) = 1$, by $g(0) > 0$ and Theorem 2, we have $\forall 0\leq x < 1$, $x < g(x)$. For the increasing property of $g(\cdot)$, so $g_n(x) < g_{n+1}(x), n\geq 1$, in addition, $g(\cdot)$ is continuous, then

$$q = \lim_{n\longrightarrow\infty} g_n(x) = \lim_{n\longrightarrow\infty} g(g_{n-1}(x)) = g\left(\lim_{n\longrightarrow\infty} g_{n-1}(x)\right), \tag{42}$$

i.e., $q = g(q)$, and then, we can get $q = 1$ by Theorem 2, i.e., $\lim_{n\longrightarrow\infty} g_n(x) = P(\infty)$.

(2) If $0 < P(\infty) < 1$, $\forall 0\leq x < P(\infty)$, then $x < g(x)$, and $g_n(x) < g_{n+1}(x), n\geq 1$; thus,

$$\begin{aligned} q &= \lim_{n\longrightarrow\infty} g_n(x) = \lim_{n\longrightarrow\infty} g(g_{n-1}(x))\\ &= g\left(\lim_{n\longrightarrow\infty} g_{n-1}(x)\right) = g(q). \end{aligned} \tag{43}$$

By Theorem 2, $q = P(\infty)$, i.e., $\lim_{n\longrightarrow\infty} g_n(x) = P(\infty)$.

(3) If $0 < P(\infty) < 1$, $\forall P(\infty)\leq x < 1$, then $x\geq g(x)$, and $g_n(x)\geq g_{n+1}(x), n\geq 1$, and thus,

$$\begin{aligned} q &= \lim_{n\longrightarrow\infty} g_n(x) = \lim_{n\longrightarrow\infty} g(g_{n-1}(x))\\ &= g\left(\lim_{n\longrightarrow\infty} g_{n-1}(x)\right) = g(q). \end{aligned} \tag{44}$$

Noting that the equation $x = g(x)$ has no root on the interval $(P(\infty), 1)$, so $q = P(\infty)$, i.e.,

$$\lim_{n\longrightarrow\infty} g_n(x) = P(\infty). \tag{45}$$

Note: the significance of Theorem 4 is obvious. It gives a numerical method to calculate the asymptotic value of extinction probability. For any initial value x_0 $(0\leq x_0 < 1)$, iteration value $g_n(x_0)$ is the asymptotic value of the extinction probability $P(\infty)$. □

4. The t-Pre-Extinction Probability

In this section, the analytic formula and the approximation algorithm of t-pre-extinction probability are given, and the iterative integral equation with unique solution is established, which is satisfied by the t-pre-extinction probability.

Let $t > 0$, $n \geq 1$, denote $\Delta_n = (t/2^n)$. Divide the interval $(0, t]$ equally into 2^n intervals $(k\Delta_n, (k+1)\Delta_n], k = 0, 1, 2, \ldots, 2^n - 1$. Step function is defined as follows:

$$H_n(s) = \begin{cases} 0, & 0 \leq s \leq \Delta_n, \\ H_n(\Delta_n), & \Delta_n < s \leq 2\Delta_n, \\ \cdots & \cdots \\ H_n(k\Delta_n), & k\Delta_n < s \leq (k+1)\Delta_n, \\ \cdots & \cdots \\ H_n(t - \Delta_n), & t - \Delta_n < s \leq t, \end{cases} \tag{46}$$

where

$$\begin{aligned} H_n(\Delta_n) &= \int_0^{\Delta_n} e^{-\int_0^s (1 - H_n(\Delta_n - u))\beta(u)\mathrm{d}u} \\ &\quad \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s; \\ H_n(k\Delta_n) &= \int_0^{k\Delta_n} e^{-\int_0^s (1 - H_n(k\Delta_n - u))\beta(u)\mathrm{d}u} \\ &\quad \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s, \quad k \end{aligned} \tag{47}$$

We always assume that the birth rate function $\beta(\cdot)$ is bounded in any finite interval; denote

$$\beta = \sup_{0 \leq u \leq t} \beta(u). \tag{48}$$

Theorem 5

(1) $\forall n \geq 1$, $H_n(\cdot)$ is nondecreasing on $[0, t]$

(2) $\{H_n(\cdot)\}_{n \geq 1}$ is a monotonic increasing sequence of functions

(3) $\forall t \geq 0$, $\lim_{n \longrightarrow \infty} H_n(t) = P(t)$

Proof

(1) To prove $H_n(\cdot)$ is a nondecreasing function on $[0, t]$ because

$$\begin{aligned} H_n(\Delta_n) &= \int_0^{\Delta_n} e^{-\int_0^s (1 - H_n(\Delta_n - u))\beta(u)\mathrm{d}u} \\ &\quad \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &\geq \int_0^{\Delta_n} \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s > 0. \end{aligned} \tag{49}$$

So, $H_n(\cdot)$ is nondecreasing on $[0, 2\Delta_n]$; suppose inductively that $H_n(\cdot)$ is nondecreasing on $[0, k\Delta_n]$, then

$$\begin{aligned} H_n(k\Delta_n) &= \int_0^{k\Delta_n} e^{-\int_0^s (1 - H_n(k\Delta_n - u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &\geq \int_0^{(k-1)\Delta_n} e^{-\int_0^s (1 - H_n(k\Delta_n - u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &\geq \int_0^{(k-1)\Delta_n} e^{-\int_0^s (1 - H_n((k-1)\Delta_n - u))\beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &= H_n((k-1)\Delta_n). \end{aligned} \tag{50}$$

Thus, $H_n(\cdot)$ is nondecreasing on $[0, (k+1)\Delta_n]$. It is deduced by mathematical induction that $H_n(\cdot)$ is nondecreasing on $[0, t]$.

(2) $\forall n \geq 1$, to prove $H_n(\cdot) \leq H_{n+1}(\cdot)$ on $[0, k\Delta_n]$, $k = 1, 2, \ldots, 2^n$.

$$\begin{aligned} &0 \leq s \leq \Delta_n, H_n(s) = 0, \\ &0 \leq s \leq \frac{\Delta_n}{2} = \Delta_{n+1}, H_{n+1}(s) = 0, \\ &\frac{\Delta_n}{2} < s \leq \Delta_n = 2\Delta_{n+1}, H_{n+1}(s) = H_{n+1}(\Delta_{n+1}) > 0. \end{aligned} \tag{51}$$

Hence, $H_n(\cdot) \le H_{n+1}(\cdot)$ on $[0, \Delta_n]$. Suppose inductively that $H_n(\cdot) \le H_{n+1}(\cdot)$ on $[0, k\Delta_n]$, by the definitions of $H_n(\cdot)$ and $H_{n+1}(\cdot)$, we have

$$\begin{aligned} H_n(s) &= H_n(k\Delta_n), \quad k\Delta_n < s \le (k+1)\Delta_n, \\ H_{n+1}(s) &= H_{n+1}(2k\Delta_{n+1}), \quad 2k\Delta_{n+1} = k\Delta_n < s \le k\Delta_n + \Delta_{n+1}, \\ H_{n+1}(s) &= H_{n+1}((2k+1)\Delta_{n+1}), \quad (2k+1)\Delta_{n+1} < s \le (k+1)\Delta_n, \end{aligned} \tag{52}$$

where

$$\begin{aligned} &H_{n+1}((2k+1)\Delta_{n+1}) \ge H_{n+1}(2k\Delta_{n+1}) = H_{n+1}(k\Delta_n) \\ &= \int_0^{k\Delta_n} e^{-\int_0^s (1 - H_{n+1}(k\Delta_n - u))\beta(u)\mathrm{d}u} \\ &\quad \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &\ge \int_0^{k\Delta_n} e^{-\int_0^s (1 - H_n(k\Delta_n - u))\beta(u)\mathrm{d}u} \\ &\quad \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s = H_n(k\Delta_n). \end{aligned} \tag{53}$$

Thus, it can be proved by mathematical induction that $H_n(s) \le H_{n+1}(s)$, $s \in [0, t]$, $n \ge 1$, i.e., $\{H_n(\cdot)\}_{n \ge 1}$ is a monotonic increasing sequence of functions.

(3) For simplicity, denote $\Delta = \Delta_n$, $H(\cdot) = H_n(\cdot)$. Because $P(\cdot)$ is uniform continuous on $[0, t]$, so $\forall \varepsilon > 0$, when n is sufficiently large, that is $\Delta = \Delta_n$ sufficiently small.

$$\begin{aligned} |P(u) - P(v)| &< \varepsilon, \\ |u - v| &< \Delta. \end{aligned} \tag{54}$$

The following conclusion can be deduced by mathematical induction:

$$0 \le P(k\Delta) - H(k\Delta) \le k\varepsilon\beta\Delta + o(\varepsilon\beta\Delta), \quad k = 1, 2, \ldots, 2^n. \tag{55}$$

It is easy to prove that $H(k\Delta) \le P(k\Delta)$, $k = 1, 2, \ldots, 2^n$.

$$\begin{aligned} P(\Delta) - H(\Delta) &= \int_0^\Delta \left[e^{-\int_0^s (1 - P(\Delta - u))\beta(u)\mathrm{d}u} - e^{-\int_0^s (1 - H(\Delta - u))\beta(u)\mathrm{d}u} \right] \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &= \int_0^\Delta \left[e^{\int_0^s (P(\Delta - u))\beta(u)\mathrm{d}u} - 1 \right] \cdot e^{-\int_0^s \beta(u)\mathrm{d}u} \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &\le \int_0^\Delta [\varepsilon\beta\Delta + 0(\varepsilon\beta\Delta)] \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &\le \varepsilon\beta\Delta + o(\varepsilon\beta\Delta), \end{aligned} \tag{56}$$

$$\begin{aligned} P(2\Delta) - H(2\Delta) &\le \int_0^{2\Delta} \left[e^{\int_0^s (P(2\Delta - u) - H(2\Delta - u))\beta(u)\mathrm{d}u} - 1 \right] \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &= \int_0^\Delta \left[e^{\int_0^s (P(2\Delta - u) - H(2\Delta - u))\beta(u)\mathrm{d}u} - 1 \right] \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\ &\quad + \int_\Delta^{2\Delta} \left[e^{\int_0^s (P(2\Delta - u) - H(2\Delta - u))\beta(u)\mathrm{d}u} - 1 \right] \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s. \end{aligned}$$

When $0 \le s \le \Delta$,

$$\begin{aligned}
&\int_0^s (P(2\Delta - u) - H(2\Delta - u))\beta(u)\mathrm{d}u \\
&\le \int_0^s (P(2\Delta - u) - P(\Delta) + P(\Delta) - H(\Delta))\beta(u)\mathrm{d}u \\
&\le \varepsilon\beta\Delta + o(\varepsilon\beta\Delta) \le 2\varepsilon\beta\Delta + o(2\varepsilon\beta\Delta).
\end{aligned} \tag{57}$$

When $\Delta \le s \le 2\Delta$,

$$\begin{aligned}
&\int_0^s (P(2\Delta - u) - H(2\Delta - u))\beta(u)\mathrm{d}u \\
&= \int_0^\Delta (P(2\Delta - u) - H(2\Delta - u))\beta(u)\mathrm{d}u + \int_\Delta^s (P(2\Delta - u) \\
&\quad - H(2\Delta - u))\beta(u)\mathrm{d}u \\
&\le \varepsilon\beta\Delta + o(\varepsilon\beta\Delta) + \int_\Delta^s (P(2\Delta - u) - P(\Delta) \\
&\quad + P(\Delta) - H(\Delta))\beta(u)\mathrm{d}u \\
&\le 2\varepsilon\beta\Delta + o(2\varepsilon\beta\Delta).
\end{aligned} \tag{58}$$

So,

$$\begin{aligned}
&P(2\Delta) - H(2\Delta) \\
&\le \int_0^\Delta (2\varepsilon\beta\Delta + o(2\varepsilon\beta\Delta)) \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s + \int_\Delta^{2\Delta} (2\varepsilon\beta\Delta + o(2\varepsilon\beta\Delta)) \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\
&= \int_0^{2\Delta} (2\varepsilon\beta\Delta + o(2\varepsilon\beta\Delta)) \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\
&\le 2\varepsilon\beta\Delta + o(2\varepsilon\beta\Delta).
\end{aligned} \tag{59}$$

Suppose inductively that $P(j\Delta) - H(j\Delta) \le j\varepsilon\beta\Delta + o(j\varepsilon\beta\Delta)$, $1 \le j \le k-1$. Then,

$$\begin{aligned}
P(k\Delta) - H(k\Delta) &\le \int_0^{k\Delta} \left[e^{\int_0^s (P(k\Delta - u) - H(k\Delta - u))\beta(u)\mathrm{d}u} - 1 \right] \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s \\
&= \sum_{j=0}^{k-1} \int_{j\Delta}^{(j+1)\Delta} \left[e^{\int_0^s (P(k\Delta - u) - H(k\Delta - u))\beta(u)\mathrm{d}u} - 1 \right] \cdot \alpha(s) e^{-\int_0^s \alpha(u)\mathrm{d}u} \mathrm{d}s.
\end{aligned} \tag{60}$$

When $(k-1)\Delta \le s \le k\Delta$,

$$\begin{aligned}&\int_0^s (P(k\Delta-u)-H(k\Delta-u))\beta(u)\mathrm{d}u\\&=\sum_{j=0}^{k-2}\int_{j\Delta}^{(j+1)\Delta}(P(k\Delta-u)-H(k\Delta-u))\beta(u)\mathrm{d}u+\int_{(k-1)\Delta}^{s}(P(k\Delta-u)-H(k\Delta-u))\beta(u)\mathrm{d}u,\end{aligned} \tag{61}$$

$\forall 0 \le j \le k-2$,

$$\begin{aligned}&\int_{j\Delta}^{(j+1)\Delta}(P(k\Delta-u)-H(k\Delta-u))\beta(u)\mathrm{d}u\\&=\int_{j\Delta}^{(j+1)\Delta}[P(k\Delta-u)-P((k-j-1)\Delta)+P((k-j-1)\Delta)-H((k-j-1)\Delta)]\beta(u)\mathrm{d}u\\&<\varepsilon\beta\Delta+o(\varepsilon\beta\Delta),\\&\int_{(k-1)\Delta}^{s}(P(k\Delta-u)-H(k\Delta-u))\beta(u)\mathrm{d}u\le\varepsilon\beta\Delta+o(\varepsilon\beta\Delta).\end{aligned} \tag{62}$$

So

$$\int_0^s (P(k\Delta-u)-H(k\Delta-u))\beta(u)\mathrm{d}u \le k\varepsilon\beta\Delta+o(k\varepsilon\beta\Delta). \tag{63}$$

It is obvious that when $s_1 \le s_2$, there is

$$\begin{aligned}&\int_0^{s_1}(P(k\Delta-u)-H(k\Delta-u))\beta(u)\mathrm{d}u\\&\quad\le\int_0^{s_2}(P(k\Delta-u)-H(k\Delta-u))\beta(u)\mathrm{d}u.\end{aligned} \tag{64}$$

So

$$\begin{aligned}&P(k\Delta)-H(k\Delta)\\&\le\sum_{j=0}^{k-1}\int_{j\Delta}^{(j+1)\Delta}[k\varepsilon\beta\Delta+o(k\varepsilon\beta\Delta)]\cdot\alpha(s)e^{-\int_0^s\alpha(u)\mathrm{d}u}\mathrm{d}s\\&=\int_0^{k\Delta}[k\varepsilon\beta\Delta+o(k\varepsilon\beta\Delta)]\cdot\alpha(s)e^{-\int_0^s\alpha(u)\mathrm{d}u}\mathrm{d}s\\&\le k\varepsilon\beta\Delta+o(k\varepsilon\beta\Delta).\end{aligned} \tag{65}$$

It is proved deductively that $0 \le P(k\Delta)-H(k\Delta) \le k\varepsilon\beta\Delta+o(k\varepsilon\beta\Delta)$, $k=1,2,\ldots,2^n$.

Especially, $0 \le P(t)-H(t) \le \varepsilon\beta t+o(\varepsilon\beta t)$.

That is,

$$\lim_{n\longrightarrow\infty} H_n(t)=P(t). \tag{66}$$

The proof is complete. □

Theorem 6. *The t-pre-extinction probability $P(t)$ is the unique solution of the iterative integral equation in Theorem 1.*

Proof. According to Theorem 1, the t-pre-extinction probability is the solution of the iterative integral equation in Theorem 1, and it is not difficult to deduce by Theorem 5 that the solution of the iterative integral equation is unique. Therefore, the t-pre-extinction probability is the unique solution of the iterative integral equation. The theorem is proved. □

5. The Stochastic Order of Extinction Moment

If $P(\infty)=1$, then the extinction moment $T(\omega)$ is a real-valued random variable, and the t-pre-extinction probability $P(t)$ is the distribution function of $T(\omega)$. In this section, we study the stochastic order of the extinction moment for different branching tree evolutions.

Theorem 7

(1) *Let $P_1(\infty)$ and $T_1(\omega)$ be, respectively, the extinction probability and extinction moment for the branching tree evolutions $\{G_t(\beta_1(\cdot),\alpha(\cdot))\}_{t\ge0}$; let $P_2(\infty)$ and $T_2(\omega)$ be, respectively, the extinction probability and extinction moment for the branching tree evolutions $\{G_t(\beta_2(\cdot),\alpha(\cdot))\}_{t\ge0}$. If $P_1(\infty)=1$, $P_2(\infty)=1$, and $\beta_1(\cdot)\le\beta_2(\cdot)$, then $T_1(\omega)$ is stochastically smaller than $T_2(\omega)$, that is $T_1(\omega)\le^{s.t.}T_2(\omega)$.*

(2) *Let $\widetilde{P}_1(\infty)$ and $\widetilde{T}_1(\omega)$ be, respectively, the extinction probability and extinction moment for the branching tree evolutions $\{G_t(\beta(\cdot),\alpha_1(\cdot))\}_{t\ge0}$; let $\widetilde{P}_2(\infty)$ and $\widetilde{T}_2(\omega)$ be, respectively, the extinction probability and extinction moment for the branching tree evolutions $\{G_t(\beta(\cdot),\alpha_2(\cdot))\}_{t\ge0}$. If $\widetilde{P}_1(\infty)=1$, $\widetilde{P}_2(\infty)=1$, and $\forall t>0$, $\int_0^t\alpha_1(s)ds\le\int_0^t\alpha_2(s)ds$, then $\widetilde{T}_2(\omega)$ is stochastically smaller than $\widetilde{T}_1(\omega)$, that is $\widetilde{T}_2(\omega)\le^{s.t.}\widetilde{T}_1(\omega)$.*

Proof

(1) Corresponding to the branching tree evolution $\{G_t(\beta_1(\cdot),\alpha(\cdot))\}_{t\geq 0}$ and $\{G_t(\beta_2(\cdot),\alpha(\cdot))\}_{t\geq 0}$, similarly to Theorem 5, define the step function series as $\{H_n^{(1)}(\cdot)\}_{n\geq 1}$ and $\{H_n^{(2)}(\cdot)\}_{n\geq 1}$, respectively.

By the hypothesis of $\beta_1(\cdot)\leq\beta_2(\cdot)$ and the definition of $H_n^{(i)}(\cdot), i=1,2$, applying the mathematical induction, it is easy to prove that $H_n^{(1)}(k\Delta_n)\geq H_n^{(2)}(k\Delta_n)$, $1\leq k\leq 2^n$, $n\geq 1$, and thus $H_n^{(1)}(\cdot)\geq H_n^{(2)}(\cdot)$. By Theorem 5,

$$\begin{aligned} P_1(t) &= P(T_1(\omega)\leq t) = \lim_{n\longrightarrow\infty} H_n^{(1)}(t), \quad t\geq 0,\\ P_2(t) &= P(T_2(\omega)\leq t) = \lim_{n\longrightarrow\infty} H_n^{(2)}(t), \quad t\geq 0. \end{aligned} \tag{67}$$

Thus, $P_1(t)\geq P_2(t)$, $t\geq 0$.

Because $T_1(\omega)\leq^{\text{s.t.}} T_2(\omega)\Leftrightarrow P_1(t)\geq P_2(t), t\geq 0$, so $T_1(\omega)\leq^{\text{s.t.}} T_2(\omega)$. Thus, (1) is proved.

(2) Corresponding to the branching tree evolution $\{G_t(\beta(\cdot),\alpha_1(\cdot))\}_{t\geq 0}$ and $\{G_t(\beta(\cdot),\alpha_2(\cdot))\}_{t\geq 0}$, similarly to Theorem 5, define the step function series as $\left\{\tilde{H}_n^{(1)}(\cdot)\right\}_{n\geq 1}$ and $\left\{\tilde{H}_n^{(2)}(\cdot)\right\}_{n\geq 1}$, respectively, and denote $F_i(t)=1-\mathrm{e}^{-\int_0^t \alpha_i(u)\mathrm{d}u}, t\geq 0, i=1,2$; then $F_1(t)\leq F_2(t)$, $0\leq t<\infty$. Denote $D_i(s)=I_{[0,k\triangle_n]}(s)\mathrm{e}^{-\int_0^s(1-\tilde{H}_n^{(i)}(k\triangle_n-u))\beta(u)\mathrm{d}u}, s\longrightarrow 0, i=1,2$; then $D_i(s)$ is a decreasing function of s. Applying the mathematical induction, we have

$$\begin{aligned} \tilde{H}_n^{(1)}(k\triangle_n) &= \int_0^{k\triangle_n} e^{-\int_0^s\left(1-\tilde{H}_n^{(i)}(k\triangle_n-u)\right)\beta(u)\mathrm{d}u}\mathrm{d}F_1(s)\\ &= \int_0^\infty D_1(s)\mathrm{d}F_1(s)\leq\int_0^\infty D_1(s)\mathrm{d}F_2(s)\\ &\leq \int_0^\infty D_2(s)\mathrm{d}F_2(s)=\tilde{H}_n^{(2)}(k\triangle_n), \quad 1\leq k\leq 2^n. \end{aligned} \tag{68}$$

Thus, $\tilde{H}_n^{(1)}(\cdot)\leq\tilde{H}_n^{(2)}(\cdot)$, and by Theorem 5,

$$\begin{aligned} \tilde{P}_1(t) &= P\left(\tilde{T}_1(\omega)\leq t\right) = \lim_{n\longrightarrow\infty}\tilde{H}_n^{(1)}(t), \quad t\geq 0,\\ \tilde{P}_2(t) &= P\left(\tilde{T}_2(\omega)\leq t\right) = \lim_{n\longrightarrow\infty}\tilde{H}_n^{(2)}(t), \quad t\geq 0. \end{aligned} \tag{69}$$

Then, $\tilde{P}_1(t)\leq\tilde{P}_2(t)$, so $\tilde{T}_1(\omega)\geq^{\text{s.t.}}\tilde{T}_2(\omega)$. Thus, (2) is proved and theorem is proved. □

6. Conclusions

This paper addresses an important problem in the field of branching process. The extinction probability and the t-pre-extinction probability are studied by constructing a branching tree evolution model in which the birth rate and the death rate are both dependent on node's age. The analytical formula and the approximation algorithm for the distribution of extinction moment are given; furthermore, the analytical formula and the approximation algorithm of extinction probability are given, and a necessary and sufficient condition of extinction with probability 1 is given.

Due to publishing constraints, only the population extinction is studied, the graph-topological properties and the age structure of nodes will be studied in subsequent papers.

Data Availability

No data were used to support the findings of this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by Shanghai Natural Science Foundation (no. 16ZR1414000).

References

[1] T. E. Harris, *The Theory of Branching Processes*, Springer-Verlag, Berlin, Germany, 1963.

[2] W. L. Smith and W. E. Wilkinson, "On branching processes in random environments," *The Annals of Mathematical Statistics*, vol. 40, no. 3, pp. 814–827, 1969.

[3] H.-X. Wang and D. Fang, "Asymptotic behaviour of population-size-dependent branching processes in Markovian random environments," *Journal of Applied Probability*, vol. 36, no. 02, pp. 611–619, 1999.

[4] H.-x. Wang, "Extinction of population-size-dependent branching processes in random environments," *Journal of Applied Probability*, vol. 36, no. 01, pp. 146–154, 1999.

[5] P. Haccou, P. Jagers, and V. A. Vatubin, "Branching processes: variation, growth, and extinction of populations," in *Cambridge Studies in Adaptive Dynamics*Cambridge University Press, Cambridge, UK, 2005.

[6] R. Bellman and T. Harris, "On age-dependent binary branching processes," *The Annals of Mathematics*, vol. 55, no. 2, pp. 280–295, 1952.

[7] J. Peter and F. C. Klebaner, "Population-size-dependent and age-dependent branching processes," *Stochastic Processes and Their Applications*, vol. 87, pp. 235–254, 2000.

[8] C. D. Greenman, "A path integral approach to age dependent branching processes," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2017, Article ID 033101, 2017.

[9] D. Kajunguri, E. B. Are, and J. W. Hargrove, "Improved estimates for extinction probabilities and times to extinction for populations of tsetse (glossing spp)," *PLoS Neglected Tropical Diseases*, vol. 13, no. 4, Article ID e0006973, 2019.

[10] D. Anna and F. Vadillo, "Extinction-time for stochastic population models," *Journal of Computational and Applied Mathematics*, vol. 295, pp. 159–169, 2016.

[11] O. Aydogmus, "On extinction time of a generalized endemic chain-binomial model," *Mathematical Biosciences*, vol. 279, pp. 38–42, 2016.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 8857417, 14 pages
https://doi.org/10.1155/2021/8857417

Research Article

An Ensemble of Adaptive Surrogate Models Based on Local Error Expectations

Huanwei Xu, Xin Zhang, Hao Li, and Ge Xiang

School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Correspondence should be addressed to Huanwei Xu; zhangxin96428@163.com

Received 1 September 2020; Revised 5 January 2021; Accepted 28 January 2021; Published 10 February 2021

Academic Editor: Guoqiang Wang

An ensemble of surrogate models with high robustness and accuracy can effectively avoid the difficult choice of surrogate model. However, most of the existing ensembles of surrogate models are constructed with static sampling methods. In this paper, we propose an ensemble of adaptive surrogate models by applying adaptive sampling strategy based on expected local errors. In the proposed method, local error expectations of the surrogate models are calculated. Then according to local error expectations, the new sample points are added within the dominating radius of the samples. Constructed by the RBF and Kriging models, the ensemble of adaptive surrogate models is proposed by combining the adaptive sampling strategy. The benchmark test functions and an application problem that deals with driving arm base of palletizing robot show that the proposed method can effectively improve the global and local prediction accuracy of the surrogate model.

1. Introduction

In the engineering design problem, computer simulation is usually applied to replace the real physics experiments. For complex engineering problems, sometimes the performance function is implicit, or due to cost and time limit, the surrogate model is often applied to approximate the real physical model. Commonly used surrogate models mainly include Kriging [1], artificial neural network [2], radial basis function (RBF) [3], support vector regression(SVR) [4], and polynomial response surface(PRS) [5].

When surrogate model is applied, how to find a suitable surrogate model is a difficult task. In order to improve the adaptability of the surrogate model, a reasonable choice is to use a linear weighted combination of different surrogate models, that is, an ensemble of surrogate models. Compared with the single surrogate model, an ensemble of surrogate models can save a lot of time wasted in screening the surrogate models. Many scholars have conducted in-depth research on it and have obtained many good achievements. Huang [6] found that the ensemble of surrogate models has higher prediction accuracy than the single surrogate model. Yan [7] proposed a new weight function construction method, which has the same accuracy as the optimal submodel and can improve the approximation of the true response distribution. Lu [8] found that the multisurrogate model has better optimization results than the single surrogate model's. Pan [9] applied the ensemble of surrogate models to the lightweight design of the car body, and the results achieved a better optimization effect. Liu [10] established the ensemble of surrogate models to solve the structure optimization of car parts. Xing [11] assigned weights to three single surrogate models by using the adaptive metropolis-Markov chain Monte Carlo method. Yin [12] compared the application of a single surrogate model and an ensemble of surrogate models in groundwater restoration design optimization problems, and the results showed that the ensemble of surrogate models is more robust. Li [13] proposed a surrogate-assisted particle swarm algorithm, which can effectively balance the global search and local search. Donncha [14] successfully used the ensemble of surrogate models to improve the forecasting system with significant effects. Ouyang [15] used the analysis of variance method to determine the weights of ensemble of

surrogate models. The comparison results show that the proposed method can not only improve the prediction performance of surrogate model, but also obtain a reliable solution. Chen [16] presented a new ensemble model which combines the advantages of global and local measures. The results show that the proposed ensemble model has satisfactory robustness and accuracy. Zhang [17] proposed a unified ensemble of surrogates with global and local measures for global metamodeling. It is concluded that the proposed model has superior accuracy while keeping comparable robustness and efficiency.

Although some progress has been made in the research of the ensemble of surrogate models, most of the current methods for constructing the ensemble of surrogate models are stationary sampling. The problem with stationary sampling is that, in order to obtain an ensemble of surrogate models that meets the accuracy requirements, the sample size must be large enough. Adaptive sampling can obtain new samples that benefit the quality of the surrogate model, which can minimize the total sample size. However, the current adaptive sampling is often applied for a single surrogate model [18–21]. Only a few scholars combine the adaptive sampling strategy with the ensemble of surrogate models [22, 23]. The remainder of this paper is organized as follows. Section 2 briefly reviews the main steps to establish the ensemble of surrogate models. In Section 3, the ensemble of surrogate models using adaptive sampling strategy based on local error expectations is described. The proposed method is verified by numerical examples and compared with the three classical ensembles of surrogate models in Section 4. Section 5 applies the proposed method to the engineering design problem of driving arm base of palletizing robot. Finally, the conclusions are given.

2. Establishment of the Ensemble of Surrogate Models

There are three main steps to establish the ensemble of surrogate models:

(1) Design of experiment: the experiment design methods are applied to determine the spatial distribution of sample points. Experiment design methods mainly include Central Composite Designs (CCDs) [24], Orthogonal Design [25], and Latin Hypercube Design (LHD) [26]. LHD is the most popular sampling method due to good spatial uniformity. The experiment design method used in this paper is also LHD.

(2) Establishment of the ensemble of surrogate models: the surrogate models can be divided into two categories. One is interpolation methods, such as RBF and Kriging. For these methods, the prediction errors of the sample points are zeroes, which has good unbiasedness. The other is the noninterpolation methods, such as PRS and SVR. The noninterpolation methods have certain fitting capabilities, but the surrogate models do not go through all sample points. Therefore, enough sample points are needed to ensure the high accuracy of the surrogate models, which has extremely high uncertainty. In view of the advantages and disadvantages of different surrogate models, the most commonly used surrogate models are the RBF model and the Kriging model. In this paper, these two surrogate models are combined to establish the ensemble of surrogate models. The expression of the ensemble of surrogate models is as follows [27]:

$$\widehat{y}_e(x) = \sum_{i=1}^{N} \omega_i \widehat{y}_i(x), \sum_{i=1}^{N} \omega_i = 1. \tag{1}$$

where $\widehat{y}_e$ is the predicted response value of the ensemble of surrogate models and N is the number of surrogate models. ω_i is the ith weight coefficient. $\widehat{y}_i$ is the predicted response value of the ith surrogate model. Generally speaking, the higher the prediction accuracy, the larger the weight coefficient of the corresponding surrogate model.

(3) Accuracy verification: accuracy verification of surrogate model mainly includes two aspects: global accuracy and local accuracy. root mean square error (RMSE) [28] and coefficient of determination (R^2) [29] are two main global accuracy evaluation methods. The corresponding expressions are as follows:

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}, \\ R^2 &= 1 - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}, \end{aligned} \tag{2}$$

where y_i is the actual response value of the ith test sample and $\widehat{y}_i$ is the predicted response value of the surrogate model of the ith test sample. $\overline{y}$ is the mean value of the actual response value, and n is the size of test sample points. For RMSE, the smaller the value, the higher the global prediction accuracy. The range of R^2 is not greater than 1. The value of R^2 can be negative if the fitting quality of the surrogate model is extremely low. The closer the value of R^2 to 1, the higher the accuracy of the global approximation of the surrogate model. Although RMSE can evaluate the prediction accuracy of the surrogate model, the magnitude of the specific problem greatly affects the value of RMSE, which is not as intuitive and easy to understand as R^2. The global accuracy evaluation method applied in this paper is the coefficient of determination R^2.

The local prediction accuracy evaluation method is maximum absolute error (MAE). The expression of MAE is as follows:

$$\mathrm{MAE} = \max\left|y_i - \widehat{y}_i\right|. \tag{3}$$

Similar to RMSE, the smaller the MAE, the higher the local prediction accuracy of the surrogate model. In this

paper, MAE is also used to evaluate the local prediction accuracy of the surrogate model.

3. The Ensemble of Adaptive Surrogate Models Based on Local Error Expectations

The existing adaptive sampling strategy of sample points is mainly for a specific surrogate model, which has poor versatility. In addition, due to the inconsistency of the existing adaptive sampling strategies, it will be very complicated to combine the ensemble of surrogate models with the adaptive sampling strategy. In this section, a universal adaptive sampling strategy based on local errors is proposed. By combining the new adaptive sampling strategy, the method to construct the ensemble of surrogate models is proposed.

3.1. Adaptive Sampling Based on Local Error Expectations. Since Kriging and RBF models usually can provide good accuracy for fitting highly nonlinear behaviors, so these two surrogate models are used in general engineering problems. At present, the most commonly used adaptive sampling method is the maximin distance approach proposed by Johnson [30]. Jin and Chen [31] made corresponding improvements and proposed the Maximin Scaled Distance Approach. In this paper, we also propose a universal adaptive sampling strategy based on the local error expectations named LEE strategy for different surrogate models and it is proposed to serve the construction of the ensemble of adaptive surrogate models. The process is shown in Figure 1.

The following are main steps of the LEE strategy:

(1) Build an initial surrogate model. First, LHD is used to obtain the initial sample points and obtain their response values. Since high accuracy is not required at the beginning of sampling, for different dimensional surrogate models, the initial number of sample points can be $5n_d$, $10n_d$, and $20n_d$ (n_d is the number of design variables).

(2) Calculate the expected value $E[\mathrm{AE}]$ of the local error. Use the existing sample points and their response values to construct a surrogate model, and use cross-validation error method (LOO-leave one method) to obtain the local error of each point. The local error of ith sample point is evaluated by the absolute error $\mathrm{AE}_i = |\hat{y}_i - y_i|$. Then the local error expectation E AE] can be obtained by the following expression:

$$E[\mathrm{AE}] = \frac{\sum_{i=1}^{n} \mathrm{AE}_i}{n}. \quad (4)$$

By using cross-validation error method, each sample point serves as a test point, and the other sample points serve as the sample points that constitute the surrogate model. When each sample point serves as the test point, it can reflect its importance for modeling and the uncertainty around the sample point's location. The absolute error AE_i can reflect the uncertainty around this location, and the expected absolute error $E[\mathrm{AE}]$ of all sample points can reflect the uncertainty of the overall sample points.

(3) Calculate the dominating radius of the sample points. Since the initial sample points determined by LHD have certain uniformity, the same radius can be set for each sample point. n sample points can divide the design space into $n-1$ part. In order to ensure that the radius of each sample point does not intersect as much as possible, we propose the concept of the dominating radius of the sample point. R_j is the dominating radius of the jth dimension

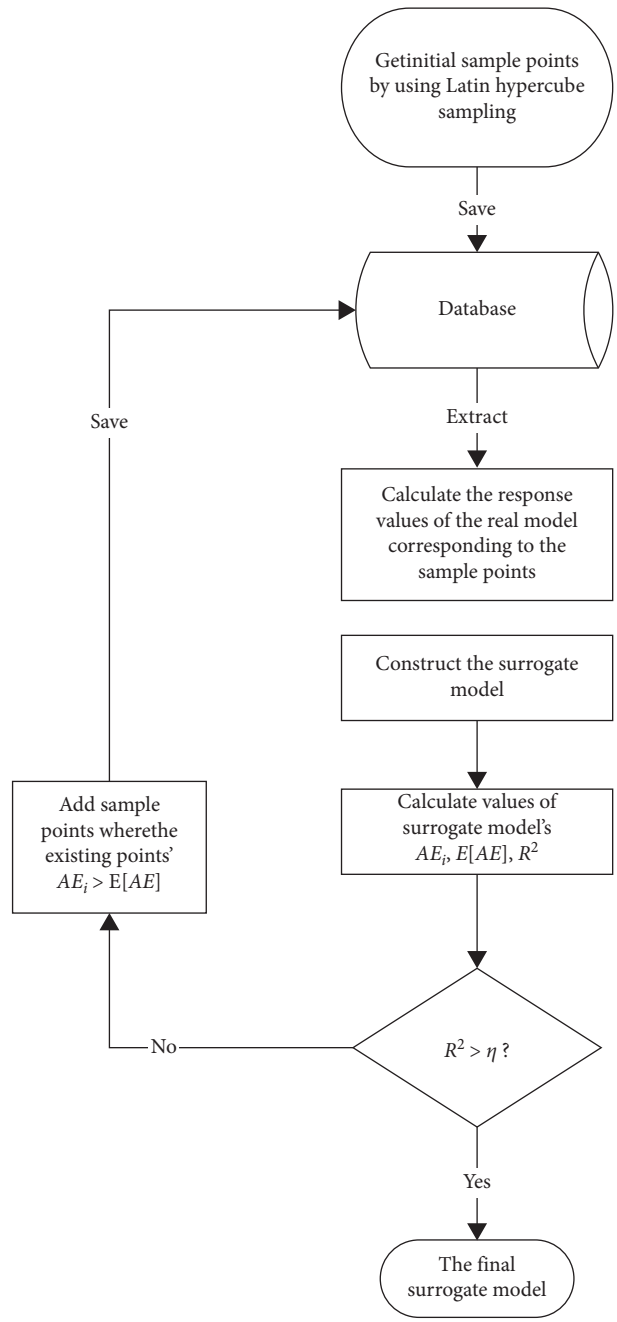


Figure 1: The adaptive sampling process based on local error expectations.

coordinate of the sample point; the expression is as follows:

$$R_j = \frac{\left|x_{j\max} - x_{j\min}\right|}{n-1}, \quad (j = 1, 2, \ldots, n_d), \tag{5}$$

where n_d is the size of the dimension and $x_{j\max}$ and $x_{j\min}$ are the upper and lower bounds of the jth dimension. Then, $R = (R_1, R_2, \ldots, R_{nd})$ is dominating radius of each sample point.

(4) Obtain new sample points. When $\mathrm{AE}_i > E[\mathrm{AE}]$, the prediction uncertainty near ith sample point is greater than the average prediction uncertainty of the existing sample points. It means the degree of nonlinearity near ith sample point is relatively large. So a sample point is randomly added within the dominating radius of ith sample point with equal probability. In order to avoid the added sample point being too close to the existing sample points, the sample point that meets the following condition is not added to the sample database:

$$\left|\left(X_{*(j)} - X_{\mathrm{closest}(j)}\right)\right| < \frac{R_j}{10}, \quad (j = 1, 2, \ldots, n_d), \tag{6}$$

where X_* stands for the point to be added and X_{closest} represents the sample point closest to point X_*. Formula (6) means that if the sample points X_* and X_{closest} are too close, they will influence the condition of the correlation matrix of the surrogate model, so the added sample point should be invalid.

(5) If the value of R^2 is greater than the preset value η, the final surrogate model is obtained; otherwise update the surrogate model. The new acquired sample points are added to the sample database. The corresponding response values of these new sample points are calculated. Then the surrogate model is updated according to the current database of sample points. Calculate the determination coefficient R^2. If the value of R^2 is greater than the preset value η, the adaptive sampling process ends; otherwise, return to step 2.

In order to illustrate the feasibility of LEE strategy, the one-dimensional test function in [32] is selected and its expression is

$$f(x) = (6x-2)^2 \sin(12x-4), \quad x \in [0, 1]. \tag{7}$$

Figures 2–4 are initial Kriging model, the absolute errors, and the updated Kriging model. Figure 2 shows that the overall prediction accuracy of the initial Kriging surrogate model is low, and the local errors near point 5 and point 6 are very large. It can be seen from Figure 3 that errors of sample points 5 and 6 of the initial Kriging model exceed $E[\mathrm{AE}]$, so random sample points are added in the dominating radius of points 5 and 6. It can be seen from Figure 4 that the added Kriging surrogate model has higher prediction accuracy. After adding the sample points, the prediction error in this area is significantly reduced, and the prediction accuracy is higher, which proves the effectiveness and feasibility of adaptive sampling based on LEE strategy.

In order to prove the versatility of LEE strategy for different surrogate models, the RBF surrogate model is also constructed based on the existing sample points and their response values. Figures 5–7 are initial RBF model, the absolute errors, and the updated RBF model. It can be seen from Figure 5 that the overall prediction accuracy of the initial RBF surrogate model is low, and the local errors near points 1 and 6 are the largest. It can be seen from Figure 6 that local errors of sample points 1 and 6 of the initial RBF model exceed $E[\mathrm{AE}]$, so random sample points are added in the dominating radius of sample points 1 and 6. It can be seen from Figure 7 that the overall prediction accuracy of updated RBF surrogate model with two new sample points has been greatly improved, which further proves the feasibility and versatility of adaptive sampling based on LEE strategy.

The proposed LEE strategy is also compared with another adaptive sampling strategy called the Maximin Scaled Distance Approach (MSDA) [31] through the classic test functions. The specific information of the test functions is shown in Table 1.

The initial Kriging and RBF surrogate models are established, respectively, according to a certain number of initial sample points. The proposed LEE strategy and MSDA are applied to improve the accuracy of surrogate models. The convergence condition is $R^2 > 0.8$. Comparison results of Kriging and RBF surrogate models are listed in Table 2.

It can be seen from Table 2 that when the numbers of initial sample points of the two methods are the same, the numbers of total sample points used by LEE strategy are less than MSDA's. At the same time, except for CN function, the final values of R^2 of the LEE strategy are greater than those of the MSDA in most functions, which means that surrogate models constructed by LEE strategy can achieve higher prediction accuracy than those constructed by MSDA.

3.2. The Ensemble of Adaptive Surrogate Models. In this section we construct the ensemble of surrogate models with LEE strategy. The flowchart is shown in Figure 8.

The main steps are as follows:

(1) Build Kriging and RBF surrogate models. Existing researches [8–12] prove that, in most cases, interpolation type (Kriging and RBF) surrogate models are more suitable for engineering problems. Therefore, this paper chooses Kriging and RBF models to form the ensemble of surrogate models. Construct Kriging and RBF models by using the initial sample points. Then, obtain the predicted error sum of square (PRESS) [33], MAE, and R^2 values of Kriging and RBF models by applying CV verification method (LOO-leave one method). The absolute errors (AEs) of each sample point of Kriging and RBF models are calculated. Since Forrester [34] has already proved that the surrogate model has better predictive ability when the coefficient of determination R^2 is greater than 0.8, we use $R^2 > 0.8$ as convergence conditions.

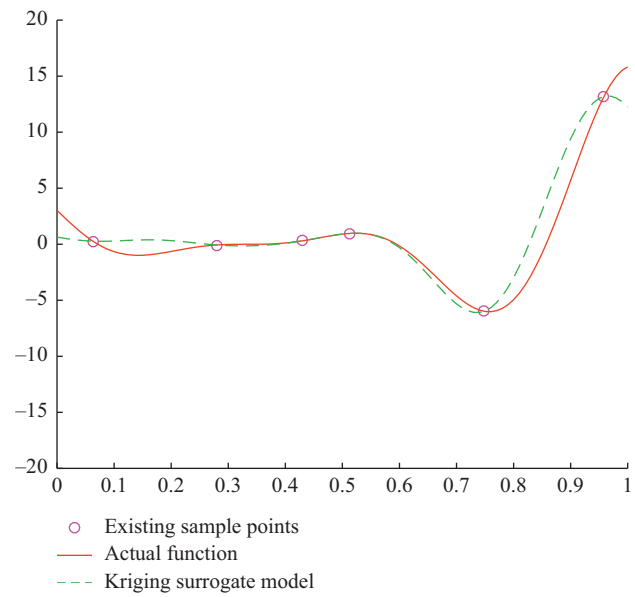


Figure 2: The initial Kriging model based on 6 sample points.

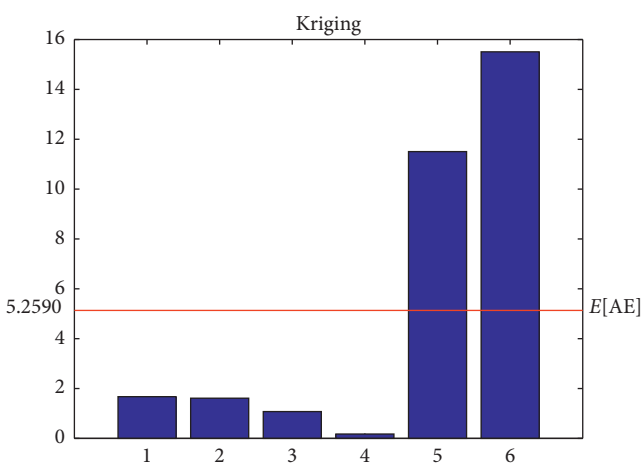


Figure 3: The absolute errors of samples of the initial Kriging model.

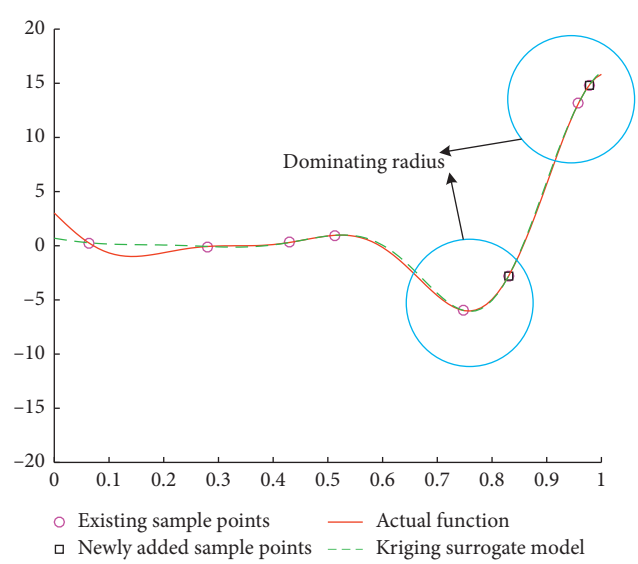


Figure 4: The updated Kriging model with 2 new sample points.

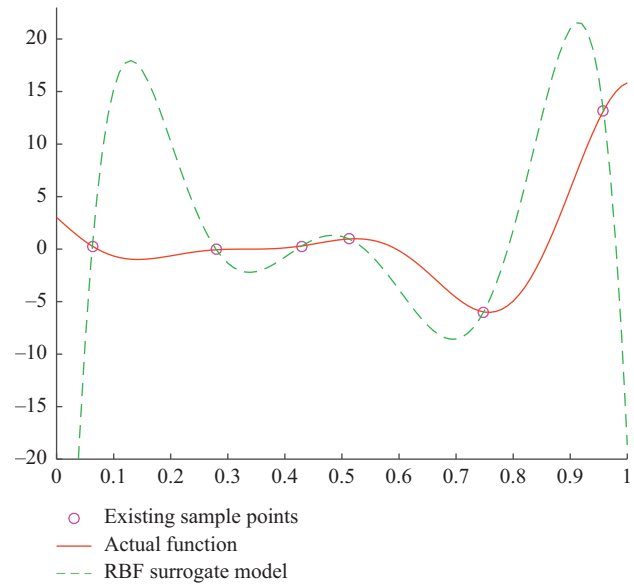


Figure 5: The initial RBF model based on 6 sample points.

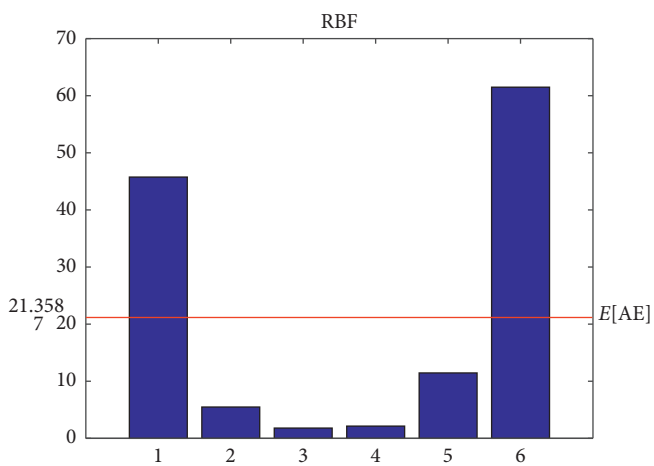


Figure 6: The absolute errors of samples of the initial RBF surrogate model.

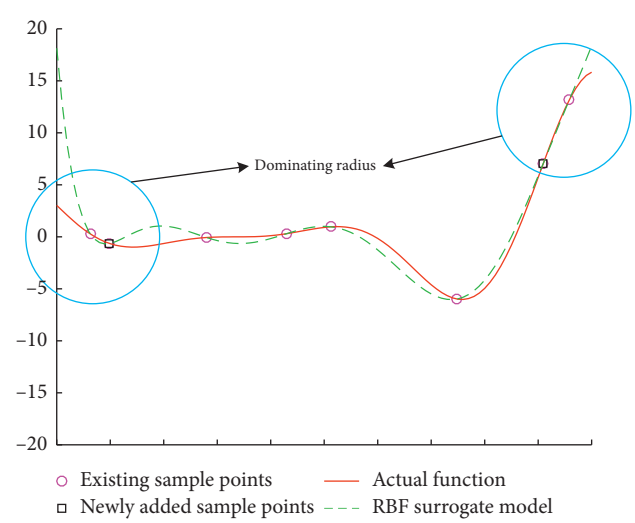


Figure 7: The updated RBF surrogate model with 2 new sample points.

Table 1: Test function expression.

Test function	Dimension	Test function expression
Branin (BN)	2	$f(x) = (x_2 - (5.1/4\pi^2)x_1^2 + (5/\pi)x_1 - 6)^2 + 10(1 - (1/8\pi))\cos(x_1) + 10$ $x_1 \in [-5, 10], x_2 \in [0, 15]$
Hartmann3 (H3)	3	$f(x) = -\sum_{i=1}^{4} \alpha_i \exp(-\sum_{j=1}^{3} A_{ij}(x_j - P_{ij})^2)$ $\alpha = (1.0, 1.2, 3.0, 3.2)^T$ $A = \begin{pmatrix} 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \end{pmatrix}$ $P = 10^{-4}\begin{pmatrix} 3689 & 1170 & 2673 \\ 4699 & 4387 & 7470 \\ 1091 & 8732 & 5547 \\ 381 & 5743 & 8828 \end{pmatrix}$ $x_i \in [0, 1]$
Colville (CV)	4	$f(x) = 100(x_1^2 - x_2)^2 + (x_1 - 1)^2 + (x_3 - 1)^2 + 90(x_3^2 - x_4)^2$ $+10.1((x_2 - 1)^2 + (x_4 - 1)^2) + 19.8(x_2 - 1)(x_4 - 1)$ $x_i \in [-10, 10], i = 1, 2$
Six-Hump Camel (SHC)	2	$f(x) = (4 - 2.1x_1^2 + (x_1^4/3))x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$ $x_1 \in [-3, 3], x_2 \in [-2, 2]$

Table 2: Comparison results of Kriging surrogate model.

Test function	Approach	The number of initial samples	Kriging model		RBF model	
			The number of total samples	Final value of R^2	The number of total samples	Final value of R^2
BN	LEE	10	18	0.946	15	0.908
	MSDA		27	0.899	27	0.873
H3	LEE	15	26	0.896	35	0.902
	MSDA		34	0.837	39	0.879
CV	LEE	20	29	0.909	30	0.934
	MSDA		44	0.943	41	0.901
SHC	LEE	10	25	0.920	21	0.941
	MSDA		36	0.883	29	0.866

(2) Obtain new sample points. The new sample points are generated by applying adaptive sampling method based on LEE strategy. The sample database is updated.

(3) Update the Kriging and the RBF models. Calculate the true response values of the newly added sample points and reconstruct the Kriging and the RBF models. As long as the R^2 of one of the two surrogate models is greater than 0.8, the operation of adding sample points is ended, and the final Kriging model and RBF model are obtained. Otherwise return to step 2.

(4) Calculate the weight coefficients of the Kriging and the RBF models and get the final ensemble of adaptive surrogate models. Cross validation (CV) [35] is performed to obtain the respective PRESS values of Kriging and RBF models. When there are n sample points in the database, all sample points except the ith point are used to construct the single surrogate model, and the ith point is used as a test point. The prediction error of the ith sample point is

$$e_i = y_i - \hat{y}_{-i}, \tag{8}$$

where y_i is the true response value of the ith sample point and $\hat{y}_{-i}$ is the predicted response value of the ith sample point in the single surrogate model composed of all sample points except ith sample point. The prediction sum of squares is the sum of the prediction errors of all sample points, as shown in the following formula:

$$\text{PRESS} = \sum_{i=1}^{n} e_i^2. \tag{9}$$

The weight coefficient corresponding to each single surrogate model is calculated by the inverse proportional

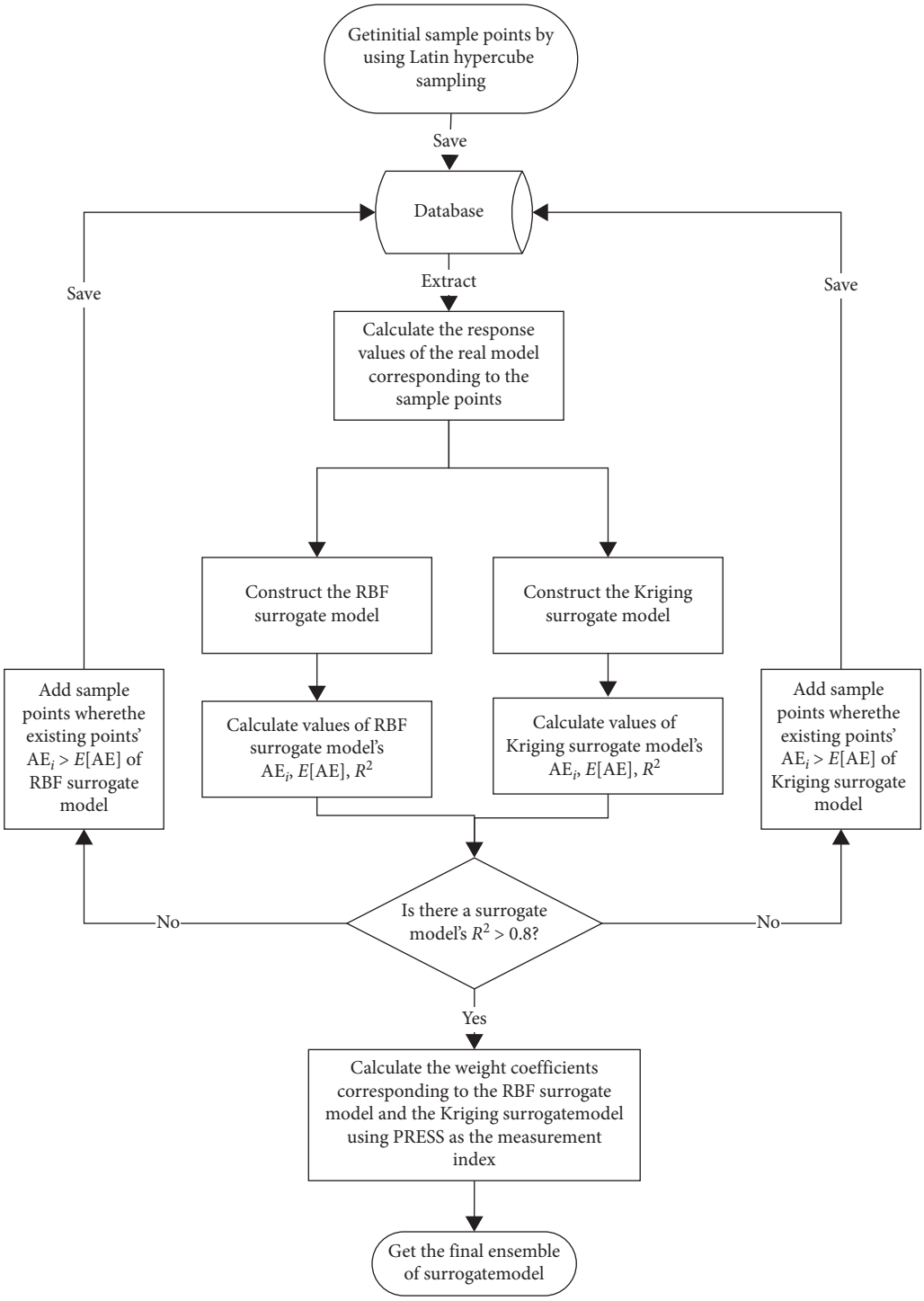


Figure 8: The construction of the ensemble of adaptive surrogate model based on LEE strategy.

averaging method, and the weight coefficient calculation formula is

$$\omega_i = \frac{(1/P_i)}{\sum_{j=1}^{N}(1/P_j)}, \tag{10}$$

where P_i is the PRESS value at the ith sample point. In this paper, N is equal to 2. Then the final ensemble of adaptive surrogate models is obtained by linearly weighting each surrogate model.

4. Numerical Example Analysis

In order to verify the versatility and effectiveness of the ensemble of adaptive surrogate models based on local error

expectations, we compare the proposed method (ensemble of adaptive surrogate model, EOASM) with three typical ensemble of surrogate model construction methods: PRESS method, BestPRESS method, and PWS (PRESS Weighted Surrogate) method [36].

Among the three most widely used methods for constructing an ensemble of surrogate model, the most classic one is to use PRESS as a measure of the weight coefficient calculation. If the PRESS value of a certain surrogate model is larger, the weight coefficient is smaller, also known as an inverse proportional averaging method, and its weight coefficient calculation formula is

$$\omega_i = \frac{(1/P_i)}{\sum_{j=1}^{N}(1/P_j)}. \tag{11}$$

The BestPRESS method selects the single surrogate model with the smallest PRESS value as the final surrogate model, which is essentially a single surrogate model. Another method is the heuristic calculation weight coefficient algorithm proposed by Goel [36], and its calculation formula is

$$\omega_i = \frac{\omega_i^*}{\sum_{j=1}^{n}\omega_j^*}, \tag{12}$$

where $\omega_i^* = (E_i + \alpha E_{\text{avg}})^{\beta}$ and $E_{\text{avg}} = (\sum_{j=1}^{n} E_j)/n$. E_i is the PRESS of the ith surrogate model. The recommended parameter values are $\alpha = 0.05, \beta = -1$.

4.1. Benchmark Functions. In this paper, six benchmark functions from low dimension to high dimension are selected. The information of benchmark functions is shown in Table 3.

The Branin, Hartmann-3, and Hartmann-4 functions are low-dimensional. Latin hypercube sampling with $5n$ sample points is enough, which meet the accuracy requirements. Since the Hartmann-6, Styblinski-Tang8, and Styblinski-Tang10 are high dimensional, the Latin hypercube sampling with $20n$ sample points is used.

4.2. The Analysis of Global Prediction Accuracy. The global prediction accuracies of different ensembles of surrogate models are compared. The total number of samples is recorded when the EOASM method reaches the convergence condition. For the other three ensembles of surrogate models constructed by the PRESS method, BestPRESS method, and PWS method, the Latin hypercube sampling method is used to generate the same total sample size. So the number of sample points in the four methods is the same. After 20 comparative experiments, the average values of the determination of coefficient R^2 of each ensemble of surrogate models are shown in Table 4.

It can be seen from Table 4 that when the total number of sample points is the same, the prediction accuracy of the ensemble of surrogate model constructed by the EOASM method is the highest. For example, for the Branin function, the average value of determination coefficient R^2 of EOASM is 0.9446. Among the other three ensembles of surrogate models, the PRESS method has the largest average value of R^2, which is much lower than that of the EOASM method. The results of the other test functions are similar to the Branin function.

4.3. The Analysis of Local Prediction Accuracy. The maximum absolute error (MAE) is used to evaluate the local accuracy. The maximum absolute error of the ensemble of surrogate model constructed by each method is compared when the number of sample points is the same. Table 5 shows the mean values of MAE of different ensembles of surrogate models.

It can be seen from 6 benchmark functions that EOASM method has the smallest average value of the MAE among four ensembles of surrogate models, which means that the proposed method has the highest predict accuracy among four methods.

4.4. Robustness Analysis. Robustness is an important indicator for evaluating surrogate models. The robustness refers to the insensitivity of the prediction accuracy of the surrogate model to random sampling of sample points. In order to compare the robustness of each surrogate model intuitively, 20 sampling experiments are performed for each benchmark function. The distribution results of the determination coefficient R^2 are presented in box plot [37], which are shown in Figure 9.

In Figure 9, the box length indicates whether the surrogate model's determination coefficient R^2 fluctuates greatly. The smaller the box length, the stronger the robustness of the surrogate model. It can be clearly seen that the box length of the ensemble of surrogate model constructed by the EOASM method is the shortest in each benchmark function, which indicates the EOASM method has the strongest robustness.

5. Engineering Application

In the design of the palletizing robot, the design of the driving arm base plays a key role. The overall assembly of the palletizing robot is shown in Figure 10.

The driving arm base bears large load. When it is assembled with the boom, it will deform to a certain extent, which will cause strain and stress. However, these physical quantities are difficult to express using explicit functions. It is often necessary to obtain their data through a large number of simulation tests. The specific material properties are shown in Table 6.

The structure of the driving arm base is shown in Figure 11. Considering the assembly relationship of each part, four nonassembly dimensions are selected as design variables, which are shown in Table 7. When the force and torque of the driving arm base reach the maximum, the generated stress is the largest. The fatigue damage is more likely to be caused. Power is carried out through UG software simulation to obtain the maximum force and torque of the assembly hole of the driving arm base.

Table 3: Test function expression.

Test function expression	Dimension	Test function expression
Branin	2	$f(x) = (x_2 - (5.1/4\pi^2)x_1^2 + (5/\pi)x_1 - 6)^2 + 10(1 - (1/8\pi))\cos(x_1) + 10$ $x_1 \in [-5, 10], x_2 \in [0, 15]$
Hartmann-3	3	$f(x) = -\sum_{i=1}^{4} \alpha_i \exp(-\sum_{j=1}^{3} A_{ij}(x_j - P_{ij})^2)$ $\alpha = (1.0, 1.2, 3.0, 3.2)^T$ $A = \begin{pmatrix} 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \end{pmatrix}$ $P = 10^{-4}\begin{pmatrix} 3689 & 1170 & 2673 \\ 4699 & 4387 & 7470 \\ 1091 & 8732 & 5547 \\ 381 & 5743 & 8828 \end{pmatrix}$ $x_i \in [0, 1]$
Hartmann-4	4	$f(x) = (1/0.839)[1.1 - \sum_{i=1}^{4} \alpha_i \exp(-\sum_{j=1}^{4} A_{ij}(x_j - P_{ij})^2)]$ $\lvert\alpha = (1.0, 1.2, 3.0, 3.2)^T$ $A = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix}$ $P = 10^{-4}\begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix}$ $x_i \in [0, 1]$
Hartmann-6	6	$f(x) = -\sum_{i=1}^{4} \alpha_i \exp(-\sum_{j=1}^{6} A_{ij}(x_j - P_{ij})^2)$ $\alpha = (1.0, 1.2, 3.0, 3.2)^T$ $A = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix}$ $P = 10^{-4}\begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix}$ $x_i \in [0, 1]$
Styblinski-Tang8	8	$f(x) = (1/2)\sum_{i=1}^{8}(x_i^4 - 16x_i^2 + 5x_i),\ x_i \in [-5, 5]$
Styblinski-Tang10	10	$f(x) = (1/2)\sum_{i=1}^{10}(x_i^4 - 16x_i^2 + 5x_i),\ x_i \in [-5, 5]$

Table 4: Mean values of R^2.

Benchmark test function	Total sample	PRESS	BestPRESS	PWS	EOASM
Branin	15	0.7351	0.7089	0.7364	0.9446
Hartmann-3	23	0.6934	0.6643	0.6935	0.9007
Hartmann-4	30	0.6549	0.5847	0.6547	0.9313
Hartmann-6	189	0.6884	0.6612	0.6883	0.9797
Styblinski-Tang8	240	0.4310	0.3903	0.4413	0.9514
Styblinski-Tang10	299	0.2931	0.2588	0.2931	0.9624

The curve of the force and torque with time is shown in Figure 12. It can be seen that, at 3 seconds, the driving arm base bears the maximum force and the maximum torque.

Since the maximum stress is difficult to calculate directly, it is selected as the object function, and its true response value is obtained by simulation with Ansys finite element software, as shown in Figure 13.

Table 5: Mean values of MAE.

Benchmark test function	Total sample	PRESS	BestPRESS	PWS	EOASM
Branin	15	108.8272	83.2561	108.7975	38.8241
Hartmann-3	23	1.2506	1.1722	1.2516	0.9772
Hartmann-4	30	2.2512	2.0696	2.2534	0.7489
Hartmann-6	189	0.4521	0.6021	0.4524	0.1578
Styblinski-Tang8	240	370.6343	331.6834	370.5411	91.3723
Styblinski-Tang10	299	330.5281	308.0860	330.7615	76.8565

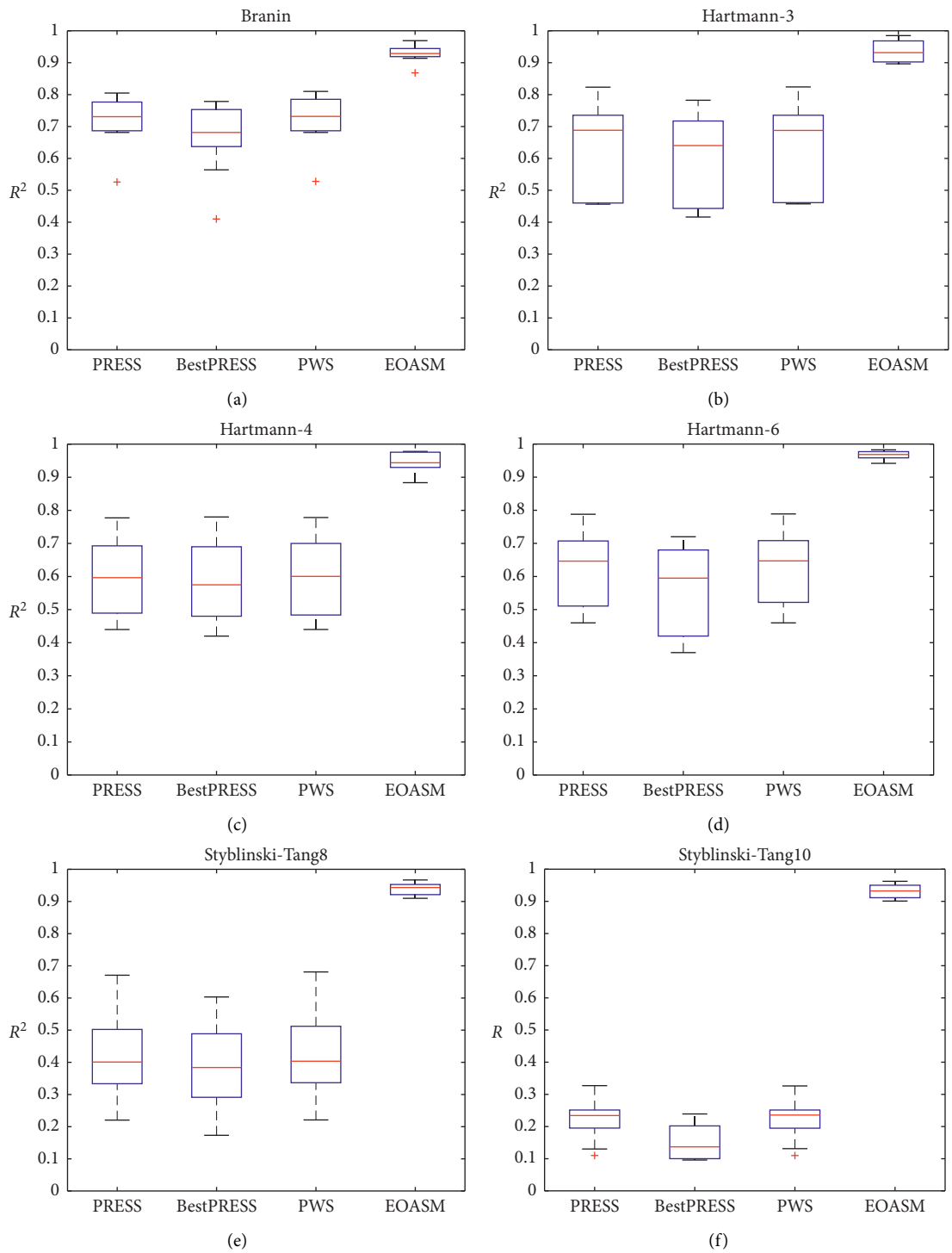


Figure 9: R^2 box diagram of the ensembles of surrogate models. (a) Branin function, (b) Hartmann-3 function, (c) Hartmann-4 function, (d) Hartmann-6 function, (e) Styblinski-Tang8 function, and (f) Styblinski-Tang10 function.

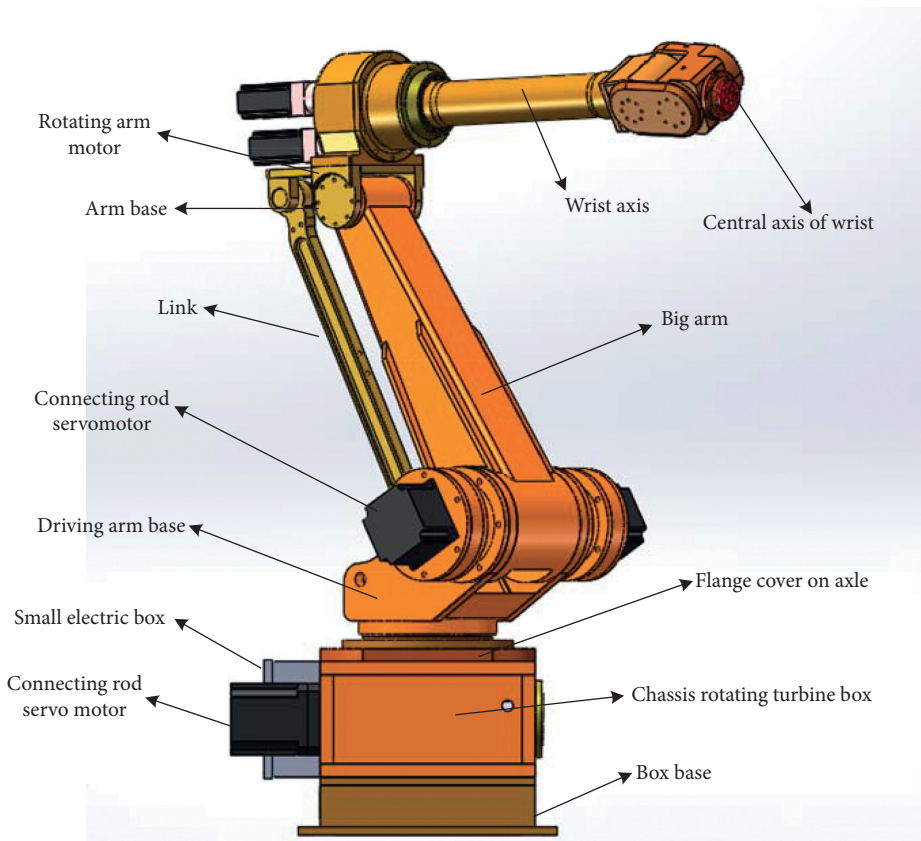


Figure 10: Overall assembly drawing of palletizing robot.

Table 6: Material properties of QT500-7.

Physical quantity	Unit	Value
Density	kg/m^3	7×10^3
Elastic modulus	Pa	1.62×10^{11}
Poisson's ratio	—	0.28
Yield strength	Pa	3.2×10^8
Tensile strength	Pa	5×10^8
Shear modulus	Pa	6.27×10^{10}

The proposed method in this paper is used to construct the ensemble of surrogate model of maximum stress. The Latin hypercube sampling is initially adopted. The number of initial sample points is $10n_d$, which is 40 sample points.

The values of global accuracy evaluation index R^2 and the local accuracy evaluation index MAE of surrogate model constructed by the EOASM method are shown in Table 8. It can be seen that the number of total sample points after convergence is 60. The CPU of the simulation platform is Intel Core i5-4590 3.30 GHz, the memory is 16G, and the operating system is Windows 10. It takes 6 minutes to perform a static structural simulation. The traditional design requires thousands of simulation experiments to roughly find the optimal value; optimization based on surrogate model only requires 60 simulation experiments, which greatly reduces computational cost of the simulation. The initial value of R^2 increases from 0.3822 to 0.8979. The global prediction accuracy is increased by 135%. Meanwhile, the value of MAE reduces from 4.1565 to 0.5007. The local

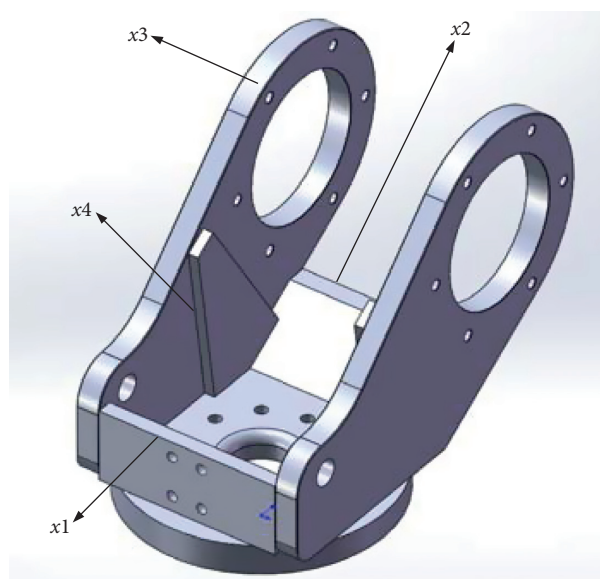


Figure 11: Driving arm base of palletizing robot.

Table 7: Design variables of driving arm base.

Design variables	Name	Unit	Ranges
x_1	Thickness of front plate	mm	13–18
x_2	Thickness of back plate	mm	8–13
x_3	Thickness of left and right board	mm	20–25
x_4	Thickness of rib	mm	8–13

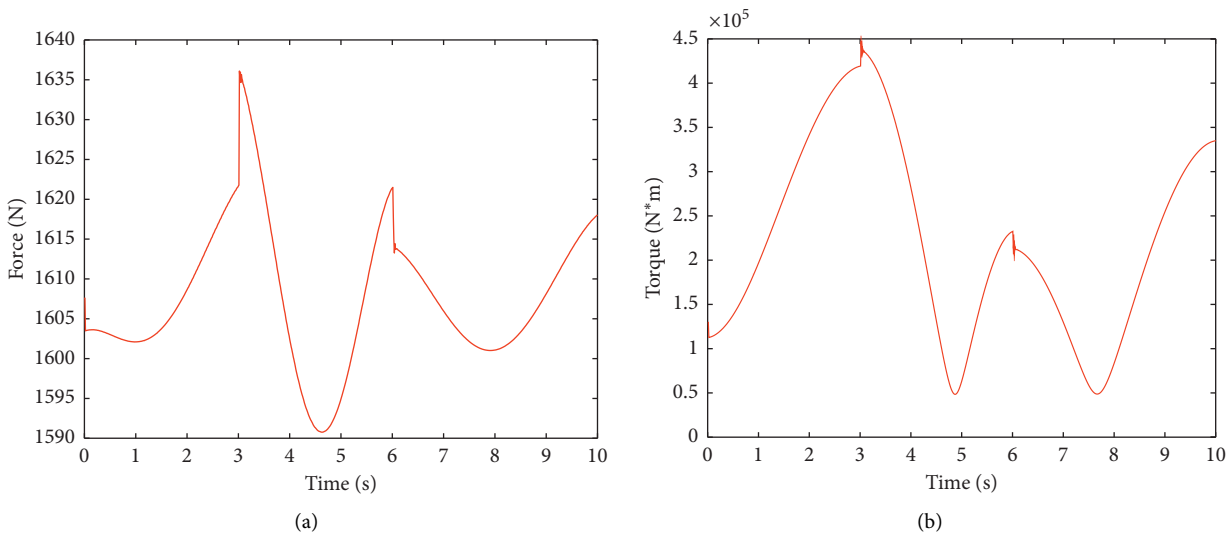


Figure 12: The force and torque of the driving arm base. (a) The force changes with time. (b) The torque changes with time.

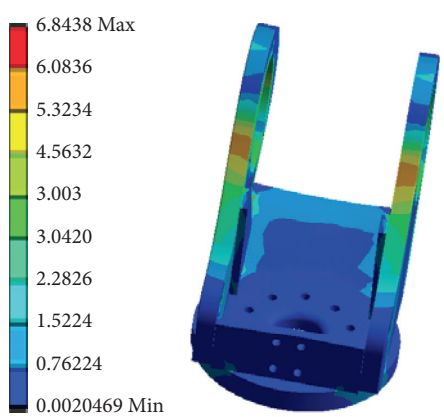


Figure 13: Stress cloud diagram of driving arm base.

Table 8: Prediction accuracy of ensemble of surrogate model constructed by EOASM method.

Evaluation perspective	The initial samples	The total samples	Initial data	EOASM data
R^2 average	40	60	0.3822	0.8979
MAE average			4.1565	0.5007

prediction accuracy is significantly improved. In summary, the EOASM method has good applicability to engineering problems and can greatly reduce the calculation cost of physical experiments.

6. Conclusion

(1) The adaptive sampling based on LEE strategy can greatly improve the prediction accuracy of the surrogate model based on as few sample points as possible, and it also has strong applicability to different types of surrogate models.

(2) The EOASM method based on LEE strategy can greatly improve the global prediction accuracy, local prediction accuracy, and the robustness of the ensemble of surrogate models.

(3) Although the prediction accuracy and robustness of the ensemble of surrogate models constructed by the EOASM method have been improved to some extent, it still has not escaped the high-dimensional curse of the surrogate model. Under the condition that the sample size is already large, it is possible that the accuracy of the surrogate model is extremely low. Therefore, the high-dimensional problem of the surrogate model is still a problem to be solved.

Data Availability

The data used to support the findings of this paper are included within the article (Table 2).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under the Contract no. 51975106.

References

[1] P. Palar and K. Shimoyama, "On efficient global optimization via universal Kriging surrogate models," *Structural & Multidisciplinary Optimization*, vol. 14, pp. 329–335, 2017.

[2] S. W. Liu, J. M. Sun, X. Gao et al., "Analysis and establishment of drilling rate prediction model based on artificial neural network," *Computer Science*, vol. 14, no. 6, pp. 605–608, 2019.

[3] S. Ozcanan and A. O. Atahan, "RBF surrogate model and EN1317 collision safety-based optimization of two guardrails," *Structural and Multidisciplinary Optimization*, vol. 60, no. 1, pp. 343–362, 2019.

[4] W. Wang, F. Dou, X. Yu et al., "Load forecasting method based on SVR under electricity market reform," *IOP Conference Series Earth and Environmental*, vol. 467, Article ID 012201, 2020.

[5] J. S. Cheng and H. Yu, "Optimal design of ossicle muffler based on response surface method," *Journal of Hunan University: Natural Science Edition*, vol. 44, no. 2, pp. 60–65, 2017.

[6] H. J. Huang, B. W. Zhang, G. Q. Wu et al., "Multidisciplinary design optimization of car body based on the ensemble of surrogate model," *Auto Motive Engineering*, vol. 38, no. 9, pp. 1107–1113, 2016.

[7] L. Yan, X. J. Duan, B. W. Liu et al., "Weighted surrogate model based on Kullback-Leibler distance dispersion," *Journal of National University of Defense Technology*, vol. 41, no. 3, pp. 159–165, 2019.

[8] Z. M. Lu, L. Q. Wang, J. Zhao et al., "An optimization method of mixed integer programming based on multi-surrogate model," *Control and Decision*, vol. 34, no. 2, pp. 362–368, 2019.

[9] F. Pan, *Research on the Ensemble of Surrogate Model Method and its Application in Lightweight Design of Car Body*, Shanghai Jiaotong University, Shanghai, China, 2011.

[10] X. Liu, *Research on Structural Optimization Design of Body Parts Based on Robust Design*, Hunan University, Hunan, China, 2013.

[11] Z. Xing, R. Qu, Y. Zhao, Q. Fu, Y. Ji, and W. Lu, "Identifying the release history of a groundwater contaminant source based on an ensemble surrogate model," *Journal of Hydrology*, vol. 572, pp. 501–516, 2019.

[12] J. Yin and T. C. F.. Tsai, "Bayesian set pair analysis and machine learning based ensemble surrogates for optimal multi-aquifer system remediation design," *Journal of Hydrology*, vol. 580, pp. 411–426, 2020.

[13] F. Li, X. Cai, and L. Gao, "Ensemble of surrogates assisted particle swarm optimization of medium scale expensive problems," *Applied Soft Computing*, vol. 580, pp. 74–79, 2018.

[14] F. O. Donncha, Y. Zhang, B. Chen et al., "Ensemble model aggregation using a computationally lightweight machine-learning model to forecast ocean waves," *Journal of Marine Systems*, vol. 199, pp. 361–378, 2019.

[15] L. Ouyang, L. Wan, C. Park, J. Wang, and Y. Ma, "Ensemble RBF modeling technique for quality design," *Journal of Management Science and Engineering*, vol. 4, no. 2, pp. 105–118, 2019.

[16] L. Chen, H. Qiu, C. Jiang, X. Cai, and L. Gao, "Ensemble of surrogates with hybrid method using global and local measures for engineering design," *Structural and Multidisciplinary Optimization*, vol. 57, no. 4, pp. 1711–1729, 2018.

[17] J. Zhang, X. Yue, J. Qiu et al., "A unified ensemble of surrogates with global and local measures for global metamodeling," *Engineering Optimization*, vol. 57, no. 1, pp. 1–22, 2020.

[18] L. Xia and D. Wang, "Aerodynamic optimization method based on Kriging adaptive surrogate model," *Aeronautical Computing Technology*, vol. 57, no. 1, pp. 17–21, 2013.

[19] N. C. Xiao, K. Yuan, and Y. S. Wang, "Structural reliability analysis method based on sequence surrogate model," *Journal of University of Electronic Science and Technology of China*, vol. 48, no. 1, pp. 156–160, 2019.

[20] Z. H. Han, Y. Zhang, C. Z. Xu et al., "Aerodynamic optimization design of large civil aircraft wings based on surrogate model," *Journal of Aeronautics*, vol. 40, no. 1, pp. 150–165, 2019.

[21] R. Mukesh, P. Soma, V. Karthikeyan et al., "Prediction of ionospheric vertical total electron content from GPS data using ordinary kriging-based surrogate model," *Astrophysics & Space Science*, vol. 364, no. 1, pp. 15–18, 2019.

[22] G. J. Zhang, *Adaptive Optimization Algorithm Based on Multi-Surrogate Model and its Application in Wheel Bearing Shaft Riveting*, South China University of Technology, Beijing, China, 2016.

[23] B. Y. Peng, *Application of Optimization Algorithm Based on Hybrid Surrogate Model in Electromagnetic Equipment Design*, Shenyang University of Technology, Beijing, China, 2019.

[24] S. M. Sanchez and P. J. Sanchez, "Very large fractional factorial and central composite designs," *ACM Transactions on Modeling and Computer Simulation*, vol. 15, no. 4, pp. 362–377, 2005.

[25] J. Wang, J. Liu, G. Zhang, J. Zhou, and K. Cen, "Orthogonal design process optimization and single factor analysis for bimodal acoustic agglomeration," *Powder Technology*, vol. 210, no. 3, pp. 315–322, 2011.

[26] D. Clifford, J. E. Payne, M. J. Pringle, R. Searle, and N. Butler, "Pragmatic soil survey design using flexible Latin hypercube sampling," *Computers & Geosciences*, vol. 67, no. 1, pp. 62–68, 2014.

[27] D. Schiavo, L. C. Trevizan, E. R. Pereira-Filho, and J. A. Nóbrega, "Evaluation of the use of multiple lines for determination of metals in water by inductively coupled plasma optical emission spectrometry with axial viewing,"

Spectrochimica Acta Part B: Atomic Spectroscopy, vol. 64, no. 6, pp. 544–548, 2009.

[28] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[29] N. J. D. Nagelkerke, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691-692, 1991.

[30] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *Journal of Statistical Planning and Inference*, vol. 26, no. 2, pp. 131–148, 1990.

[31] R. Jin, W. Chen, and A. Sudjitanto, "On adaptive sampling for global metamodeling in engineering design," 2002.

[32] A. I. J. Forrester and A. J. Keane, "Recent advances in surrogate-based optimization," *Progress in Aerospace Sciences*, vol. 45, no. 1–3, pp. 50–79, 2009.

[33] L. E. Zerpa, N. V. Queipo, S. Pintos, and J.-L. Salager, "An optimization methodology of alkaline-surfactant-polymer flooding processes using field scale numerical simulation and multiple surrogates," *Journal of Petroleum Science and Engineering*, vol. 47, no. 3-4, pp. 197–208, 2005.

[34] A. I. J. Forrester, A. Sobester, and A. J. Keane, *Engineering Design via Surrogate Modelling: A Practical Guide*, DBLP, Beijing, China, 2008.

[35] Y. J. Li, J. Zhang, Y. Y. Cao et al., "Forecasting of aero-engine performance trend based on fuzzy information granulation and optimized SVM," *Journal of Aeronautical Dynamics*, vol. 32, no. 12, pp. 3022–3030, 2017.

[36] T. Goel, R. T. Haftka, W. Shyy, and N. V. Queipo, "Ensemble of surrogates," *Structural and Multidisciplinary Optimization*, vol. 33, no. 3, pp. 199–216, 2007.

[37] F. A. C. Viana, R. T. Haftka, and L. T. Watson, "Efficient global optimization algorithm assisted by multiple surrogate techniques," *Journal of Global Optimization*, vol. 56, no. 2, pp. 669–689, 2013.

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 6674520, 11 pages
https://doi.org/10.1155/2021/6674520

Research Article

Jacobian Consistency of a Smoothing Function for the Weighted Second-Order Cone Complementarity Problem

Wenli Liu,[1] Xiaoni Chi,[2] Qili Yang,[1] and Ranran Cui[1]

[1]*School of Mathematics and Computing Science, Guangxi Key Laboratory of Cryptography and Information Security, Guilin University of Electronic Technology, Guilin 541004, China*
[2]*School of Mathematics and Computing Science, Guangxi Key Laboratory of Automatic Detecting Technology and Instruments, Guilin University of Electronic Technology, Guilin 541004, China*

Correspondence should be addressed to Xiaoni Chi; chixiaoni@126.com

Received 12 November 2020; Revised 14 December 2020; Accepted 4 January 2021; Published 23 January 2021

Academic Editor: Guoqiang Wang

In this paper, a weighted second-order cone (SOC) complementarity function and its smoothing function are presented. Then, we derive the computable formula for the Jacobian of the smoothing function and show its Jacobian consistency. Also, we estimate the distance between the subgradient of the weighted SOC complementarity function and the gradient of its smoothing function. These results will be critical to achieve the rapid convergence of smoothing methods for weighted SOC complementarity problems.

1. Introduction

The weighted second-order cone complementarity problem (WSOCCP) is, for a given weight vector $w \in \mathcal{K}$ and a continuously differentiable function $F\colon \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^{n+m}$, to find vectors $(x, s, y) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$ such that

$$\begin{cases} x \circ s = w, \\ F(x, s, y) = 0, \\ x \in \mathcal{K}, \\ s \in \mathcal{K}, \end{cases} \tag{1}$$

where $\circ$ represents the Jordan product and $\mathcal{K}$ is the Cartesian product of second-order cone, that is, $\mathcal{K} = \mathcal{K}^{n_1} \times \mathcal{K}^{n_2} \times \cdots \times \mathcal{K}^{n_r}$ with $\sum_{i=1}^{r} n_i = n, i = 1, \ldots, r$. The set $\mathcal{K}^{n_i}\,(i = 1, \ldots, r)$ is the second-order cone (SOC) of dimension n_i defined by

$$\mathcal{K}^{n_i} := \left\{x_i = (x_{i0}, x_{i1}) \in \mathbb{R} \times \mathbb{R}^{n_i - 1}\colon x_{i0} - \|x_{i1}\| \geq 0\right\}, \tag{2}$$

and the interior of the SOC $\mathcal{K}^{n_i}$ is the set

$$\operatorname{int}\mathcal{K}^{n_i} = \left\{x_i = (x_{i0}, x_{i1}) \in \mathbb{R} \times \mathbb{R}^{n_i - 1}\colon x_{i0} - \|x_{i1}\| > 0\right\}. \tag{3}$$

Here $\|\cdot\|$ is the Euclidean norm, and

$$\operatorname{int}\mathcal{K} = \operatorname{int}\mathcal{K}^{n_1} \times \operatorname{int}\mathcal{K}^{n_2} \times \cdots \times \operatorname{int}\mathcal{K}^{n_r}. \tag{4}$$

Obviously, if $w = 0$, WSOCCP (1) reduces to second-order cone complementarity problem (SOCCP). In this article, we may assume that $r = 1$ and $\mathcal{K} = \mathcal{K}^n$ in the following analysis, since it can easily be extended to the general case.

In order to reformulate several equilibrium problems in economics and study highly efficient algorithms to solve these problems, Potra [1] introduced the notion of a weighted complementarity problem (WCP). He showed that the Fisher market equilibrium problem can be modeled as a monotone linear WCP. Moreover, the linear programming and weighted centering (LPWC) problem, which was introduced by Anstreicher [2], can also be formulated as a monotone linear WCP. And Potra [1] analyzed two interior-point methods for solving the monotone linear WCP over the nonnegative orthant. Since then, many scholars are dedicated to investigating the theories and solution methods

of WCP. Tang [3] gave a new nonmonotone smoothing-type algorithm to solve the linear WCP. Chi et al. [4] studied the existence and uniqueness of the solution for a class of WCPs.

As is well known, smoothing methods have superior theoretical and numerical performances. For solving the SOCCP by smoothing methods, we usually reformulate the SOCCP as a system of equations based on parametric smoothing functions of SOC complementarity functions [5, 6]. The smoothing parameter involved in smoothing functions may be treated as a variable [7] or a parameter with an appropriate parameter control [8]. In the latter case, the Jacobian consistency is important to achieve a rapid convergence of Newton methods or Newton-like methods. Hayashi et al. [8] proposed a combined smoothing and regularized method for monotone SOCCP, and based on the Jacobian consistency of the smoothing natural residual function, they proved that the method has global and quadratic convergence. Krejić and Rapajić [9] gave a nonmonotone Jacobian smoothing inexact Newton method for nonlinear complementarity problem and proved the global and local superlinear convergence of the method. Chen et al. [10] presented a modified Jacobian smoothing method for the nonsmooth complementarity problem and established the global and fast local convergence for the method.

In this paper, we consider the function φ: $\mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ for WSOCCP

$$\varphi(x, s, w) := x + s - \sqrt{x^2 + s^2 + x \circ s + w}, \tag{5}$$

with a given vector $w \in \mathcal{K}^n$. If $w = 0$, φ (5) reduces to the SOC complementarity function [6] with $\tau = 3$:

$$\varphi(x, s, 0) := x + s - \sqrt{x^2 + s^2 + x \circ s}. \tag{6}$$

Since φ is nonsmooth, we define the following smoothing function φ_μ:

$$\varphi_\mu(x, s, w) := x + s - \sqrt{x^2 + s^2 + x \circ s + w + \mu^2 e}, \tag{7}$$

where $\mu \in \mathbb{R}$ is a smoothing parameter.

The main contribution of this paper is to show the Jacobian consistency of the smoothing function (7) and estimate the distance between the subgradient of the weighted SOC complementarity function (5) and the gradient of its smoothing function (7). These properties will be critical to solve weighted SOC complementarity problems by smoothing methods.

The paper is organized as follows. In Section 2, we review some concepts and properties. In Section 3, we derive the computable formula for the Jacobian of the smoothing function in WSOCCP. In Section 4, we show the Jacobian consistency of the smoothing function and estimate the distance between the gradient of smoothing function and the subgradient of the weighted SOC complementarity function. Some conclusions are reported in Section 5.

Throughout this paper, $\mathbb{R}_+$ denotes the set of nonnegative numbers. $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ denote the space of n-dimensional real column vectors and the space of matrices, respectively. We use $\|\cdot\|$ to denote the Euclidean norm and define $\|x\| := \sqrt{x^T x}$ for a vector x or the corresponding induced matrix norm. For simplicity, we often use $x = (x_0; x_1)$ instead of the column vector $x = (x_0, x_1^T)^T$. $\operatorname{int}\mathcal{K}^n$ and $\operatorname{bd}\mathcal{K}^n$ mean the topological interior and the boundary of the SOC $\mathcal{K}^n$, respectively. For a given set $S \subset \mathbb{R}^{m\times n}$, $\operatorname{conv} S$ denotes the convex hull of S in $\mathbb{R}^{m\times n}$, and for any matrix $X \in \mathbb{R}^{m\times n}$, $\operatorname{dist}(X, S)$ denotes $\inf\{\|X - Y\|: Y \in S\}$.

2. Preliminaries

In this section, we briefly recall some definitions and results about the Euclidean Jordan algebra [11] associated with the SOC $\mathcal{K}^n$ and subdifferentials [12].

For any $x, s \in \mathbb{R}^n$, their Jordan product is defined as $x \circ s = (x^T s; x_0 s_1 + s_0 x_1)$, and $e = (1, 0, \ldots, 0) \in \mathbb{R}^n$ is unit element of this algebra. Given an element $x = (x_0; x_1) \in \mathbb{R} \times \mathbb{R}^{n-1}$, we define the symmetric matrix

$$L(x) = \begin{pmatrix} x_0 & x_1^T \\ x_1 & x_0 I \end{pmatrix}, \tag{8}$$

where I represents the $(n-1) \times (n-1)$ identity matrix. It is easy to verify that $x \circ s = L(x)s$ for any $s \in \mathbb{R}^n$. Moreover, $L(x)$ is positive definite (and hence invertible) if and only if $x \in \operatorname{int}\mathcal{K}^n$.

For each $x = (x_0; x_1) \in \mathbb{R} \times \mathbb{R}^{n-1}$, let λ_1, λ_2 and $u^{(1)}, u^{(2)}$ be the spectral values and the associated spectral vectors of x, given by

$$\lambda_i = x_0 + (-1)^i \|x_1\|,$$
$$u^{(i)} = \begin{cases} \dfrac{1}{2}\left(1; (-1)^i \dfrac{x_1}{\|x_1\|}\right), & \text{if } x_1 \neq 0, \\ \dfrac{1}{2}\left(1; (-1)^i \overline{x}_1\right), & \text{otherwise}, \end{cases} \tag{9}$$

for $i = 1, 2$, with any $\overline{x}_1 \in \mathbb{R}^{n-1}$ such that $\|\overline{x}_1\| = 1$. Then, x admits a spectral factorization associated with SOC $\mathcal{K}^n$ in the form of

$$x = \lambda_1 u^{(1)} + \lambda_2 u^{(2)}. \tag{10}$$

For any $x = (x_0; x_1) \in \mathbb{R} \times \mathbb{R}^{n-1}$, let $x' = (x_0; -x_1)$[13]. Then, $x'' = x$, $(x + s)' = x' + s'$, and $(cx)' = cx'$ for any $c \in \mathbb{R}$. Moreover, $x \circ x' = x_0^2 - \|x_1\|^2 = 0$ if $x \in \operatorname{bd}\mathcal{K}^n$.

Suppose that G: $\mathbb{R}^m \longrightarrow \mathbb{R}^n$ is a locally Lipschitzian function; then, from Rademacher's theorem [14], G is differentiable almost everywhere. The Bouligand (B-) subdifferential and the Clarke subdifferential of G at z are defined by

$$\partial_B G(z) := \left\{\lim_{\widehat{z} \longrightarrow z} G'(\widehat{z}): \widehat{z} \in D_G\right\} \text{ and } \partial G(z) = \operatorname{conv}\partial_B G(z), \tag{11}$$

where D_G denotes the set of points at which G is differentiable. Obviously, $\partial G(z) = \{G'(z)\}$ if G is continuously differentiable at z.

Definition 1 (see [12]). Let $G\colon \mathbb{R}^m \longrightarrow \mathbb{R}^n$ be a locally Lipschitzian function and $G_\mu\colon \mathbb{R}^m \longrightarrow \mathbb{R}^n$ be a continuously differentiable function for any $\mu > 0$, and for any $z \in \mathbb{R}^m$, we have $\lim_{\mu\longrightarrow 0} G_\mu(z) = G(z)$. Then, G_μ satisfies the Jacobian consistency property if for any $z \in \mathbb{R}^m$, $\lim_{\mu\longrightarrow 0}\operatorname{dist}(G'_\mu(z), \partial G(z)) = 0$.

3. Smoothing Function

In this section, we study the properties of the smoothing function (7).

Definition 2 (see [8]). For a nondifferentiable function $f\colon \mathbb{R}^m \longrightarrow \mathbb{R}^n$, we consider a function $f_\mu\colon \mathbb{R}^m \longrightarrow \mathbb{R}^n$ with a parameter $\mu > 0$ that has the following properties:

(i) f_μ is differentiable for any $\mu > 0$

(ii) $\lim_{\mu\longrightarrow 0} f_\mu(x) = f(x)$ for any $x \in \mathbb{R}^m$

Such a function f_μ is called a smoothing function of f.

Lemma 1. *For any $w \in \mathcal{K}^n$ and $\mu \in \mathbb{R}$, one has*

$$\varphi_\mu(x,s,w) = 0 \Leftrightarrow x \circ s = w + \mu^2 e, \quad x \in \mathcal{K}^n, s \in \mathcal{K}^n. \tag{12}$$

Proof. We first suppose that $x \circ s = w + \mu^2 e, x \in \mathcal{K}^n, s \in \mathcal{K}^n$. Then,

$$\begin{aligned} 0 &= x \circ s - w - \mu^2 e \\ &= (x+s)^2 - \left(x^2 + s^2 + x \circ s + w + \mu^2 e\right), \end{aligned} \tag{13}$$

and hence

$$x + s = \sqrt{x^2 + s^2 + x \circ s + w + \mu^2 e}. \tag{14}$$

That is, $\varphi_\mu(x,s,w) = 0$.

Conversely, suppose that $\varphi_\mu(x,s,w) = 0$; then, it follows from (7) that

$$x + s = \sqrt{x^2 + s^2 + x \circ s + w + \mu^2 e} \in \mathcal{K}^n. \tag{15}$$

Upon squaring both sides of it, we obtain

$$x \circ s = w + \mu^2 e \in \mathcal{K}^n. \tag{16}$$

Let

$$\omega := x + s = \sqrt{x^2 + s^2 + x \circ s + w + \mu^2 e} \in \mathcal{K}^n, \tag{17}$$

which implies

$$\begin{aligned} \omega &\in \mathcal{K}^n, \\ \omega^2 &= x^2 + s^2 + x \circ s + w + \mu^2 e \in \mathcal{K}^n. \end{aligned} \tag{18}$$

Therefore,

$$\begin{aligned} \omega^2 - s^2 &= x^2 + x \circ s + w + \mu^2 e \in \mathcal{K}^n, \\ \omega^2 - x^2 &= s^2 + x \circ s + w + \mu^2 e \in \mathcal{K}^n. \end{aligned} \tag{19}$$

Further, it follows from Proposition 3.4 [15] that

$$\begin{aligned} x &= \omega - s \in \mathcal{K}^n, \\ s &= \omega - x \in \mathcal{K}^n. \end{aligned} \tag{20}$$

□

Let $w = (w_0; w_1) \in \mathcal{K}^n$, $\mu \in \mathbb{R}$, $x = (x_0; x_1)$, $s = (s_0; s_1) \in \mathbb{R} \times \mathbb{R}^{n-1}$, and the mapping $v^\mu\colon \mathbb{R}^{2n} \longrightarrow \mathbb{R} \times \mathbb{R}^{n-1}$ be defined by

$$v^\mu = \left(v_0^\mu; v_1^\mu\right) = v^\mu(x,s,w) := x^2 + s^2 + x \circ s + w + \mu^2 e, \tag{21}$$

For simplicity, we use v to denote v^μ when $\mu = 0$, that is,

$$v = (v_0; v_1) = v(x,s,w) := x^2 + s^2 + x \circ s + w. \tag{22}$$

By direct calculations, we have

$$\begin{aligned} v_0^\mu &= \|x\|^2 + \|s\|^2 + x^T s + w_0 + \mu^2 = v_0 + \mu^2, \\ v_1^\mu &= 2x_0 x_1 + 2s_0 s_1 + x_0 s_1 + s_0 x_1 + w_1 = v_1. \end{aligned} \tag{23}$$

Therefore, $v^\mu = (v_0^\mu; v_1)$. From the definition of spectral factorization, v^μ can be decomposed as

$$v^\mu = \lambda_1(v^\mu) u_1(v) + \lambda_2(v^\mu) u_2(v), \tag{24}$$

where $\lambda_1(v^\mu), \lambda_2(v^\mu)$, and $u_1(v), u_2(v)$ are the spectral values and the associated spectral vectors of v^μ given by

$$\begin{aligned} \lambda_i(v^\mu) &= \|x\|^2 + \|s\|^2 + x^T s + w_0 + \mu^2 \\ &\quad + (-1)^i \left\| 2x_0 x_1 + 2s_0 s_1 + x_0 s_1 + s_0 x_1 + w_1 \right\|, \end{aligned} \tag{25}$$

and

$$u_i(v) = \frac{1}{2}\left(1; (-1)^i \overline{v}_1\right), \tag{26}$$

for $i = 1, 2$, where

$$\overline{v}_1 := \frac{v_1}{\|v_1\|} = \frac{2x_0 x_1 + 2s_0 s_1 + x_0 s_1 + s_0 x_1 + w_1}{\left\| 2x_0 x_1 + 2s_0 s_1 + x_0 s_1 + s_0 x_1 + w_1 \right\|}, \tag{27}$$

if $v_1 \neq 0$; otherwise, $\overline{v}_1$ is any vector in $\mathbb{R}^{n-1}$ such that $\|\overline{v}_1\| = 1$. For any given $w = (w_0; w_1) \in \mathcal{K}^n$ and any $(x,s) \in \mathbb{R}^n \times \mathbb{R}^n$, it can be verified that

$$\begin{aligned} v^\mu &= x^2 + s^2 + x \circ s + w + \mu^2 e \\ &= \left(x + \frac{s}{2}\right)^2 + \frac{3}{4}s^2 + w + \mu^2 e \\ &= \left(s + \frac{x}{2}\right)^2 + \frac{3}{4}x^2 + w + \mu^2 e \in \operatorname{int}\mathcal{K}^n, \end{aligned} \tag{28}$$

for any $\mu > 0$, and

$$\begin{aligned} v &= x^2 + s^2 + x \circ s + w \\ &= \left(x + \frac{s}{2}\right)^2 + \frac{3}{4}s^2 + w \\ &= \left(s + \frac{x}{2}\right)^2 + \frac{3}{4}x^2 + w \in \mathcal{K}^n. \end{aligned} \tag{29}$$

Given $\mu \in \mathbb{R}$ and $x = (x_0; x_1), s = (s_0; s_1) \in \mathbb{R} \times \mathbb{R}^{n-1}$, we define

$$\omega^\mu = (\omega_0^\mu; \omega_1^\mu) = \omega^\mu(x, s, w) := \sqrt{x^2 + s^2 + x \circ s + w + \mu^2 e}, \tag{30}$$

and when $\mu = 0$,

$$\omega = (\omega_0; \omega_1) = \omega(x, s, w) := \sqrt{x^2 + s^2 + x \circ s + w}. \tag{31}$$

The spectral factorization of ω^μ and ω is as follows:

$$\begin{aligned} \omega^\mu &= \sqrt{\lambda_1(v^\mu)} u_1(v) + \sqrt{\lambda_2(v^\mu)} u_2(v), \\ \omega &= \sqrt{\lambda_1(v)} u_1(v) + \sqrt{\lambda_2(v)} u_2(v). \end{aligned} \tag{32}$$

By (29), we can partition $\mathbb{R}^{2n}$ as $\mathbb{R}^{2n} = \mathcal{O} \cup \mathcal{I} \cup \mathcal{B}$, where

$$\begin{aligned} \mathcal{O} &:= \left\{(x, s) \in \mathbb{R}^{2n}: v \in \{0\}\right\} \\ &= \left\{(x, s) \in \mathbb{R}^{2n}: \lambda_2(v) = \lambda_1(v) = 0\right\}, \\ \mathcal{I} &:= \left\{(x, s) \in \mathbb{R}^{2n}: v \in \operatorname{int}\mathcal{K}^n\right\} \\ &= \left\{(x, s) \in \mathbb{R}^{2n}: \lambda_2(v) \geq \lambda_1(v) > 0\right\}, \\ \mathcal{B} &:= \left\{(x, s) \in \mathbb{R}^{2n}: v \in bd\mathcal{K}^n / \{0\}\right\} \\ &= \left\{(x, s) \in \mathbb{R}^{2n}: 2v_0 = \lambda_2(v) > \lambda_1(v) = 0\right\}. \end{aligned} \tag{33}$$

Lemma 2. *For any given $w \in \mathcal{K}^n$ and any $(\mu, x, s) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$, let φ and φ_μ be defined as (5) and (7), respectively. Then, we have*

(i) The function φ_μ is continuously differentiable everywhere with any $\mu > 0$, and its Jacobian is given by

$$\varphi_\mu'(x, s, w) = \begin{pmatrix} I - L\left(x + \frac{s}{2}\right) L^{-1}(\omega^\mu) \\ \\ I - L\left(s + \frac{x}{2}\right) L^{-1}(\omega^\mu) \end{pmatrix}. \tag{34}$$

Here $L^{-1}(\omega^\mu) = (1/\sqrt{v_0^\mu}) I$ if $v_1 = 0$; otherwise,

$$\begin{aligned} L^{-1}(\omega^\mu) &= L_1(v^\mu) + L_2(v^\mu) \\ &= \begin{pmatrix} b_\mu & c_\mu \overline{v}_1^T \\ c_\mu \overline{v}_1 & a_\mu I + (b_\mu - a_\mu) \overline{v}_1 \overline{v}_1^T \end{pmatrix}, \end{aligned} \tag{35}$$

with

$$L_1(v^\mu) = \frac{1}{2\sqrt{\lambda_1(v^\mu)}} \begin{pmatrix} 1 & -\overline{v}_1^T \\ -\overline{v}_1 & \overline{v}_1 \overline{v}_1^T \end{pmatrix}, \tag{36}$$

$$L_2(v^\mu) = \frac{1}{2\sqrt{\lambda_2(v^\mu)}} \begin{pmatrix} 1 & \overline{v}_1^T \\ \overline{v}_1 & \overline{v}_1 \overline{v}_1^T \end{pmatrix} + a_\mu \begin{pmatrix} 0 & 0^T \\ 0 & I - \overline{v}_1 \overline{v}_1^T \end{pmatrix}, \tag{37}$$

where

$$\begin{aligned} a_\mu &= \frac{2}{\sqrt{\lambda_1(v^\mu)} + \sqrt{\lambda_2(v^\mu)}}, \\ b_\mu &= \frac{1}{2}\left(\frac{1}{\sqrt{\lambda_1(v^\mu)}} + \frac{1}{\sqrt{\lambda_2(v^\mu)}}\right), \\ c_\mu &= \frac{1}{2}\left(\frac{1}{\sqrt{\lambda_2(v^\mu)}} - \frac{1}{\sqrt{\lambda_1(v^\mu)}}\right). \end{aligned} \tag{38}$$

(ii) For any $(x, s) \in \mathbb{R}^n \times \mathbb{R}^n$, we have $\lim_{\mu \longrightarrow 0} \varphi_\mu(x, s, w) = \varphi(x, s, w)$. Thus, φ_μ is a smoothing function of φ.

(iii) For any $\mu, \nu \in \mathbb{R}_+$,

$$\left\|\varphi_\mu(x, s, w) - \varphi_\nu(x, s, w)\right\| \leq \sqrt{r}|\mu - \nu|. \tag{39}$$

Proof

(i) For any $(x, s) \in \mathbb{R}^n \times \mathbb{R}^n$ and any $\mu > 0$, according to Corollary 5.4 [15] and (28), formula (34) holds. By Proposition 5.2 and its proof [15], we get formula (35).

(ii) Given any $x = (x_0; x_1), s = (s_0; s_1) \in \mathbb{R} \times \mathbb{R}^{n-1}$. For any $\mu > 0$, we obtain from the spectral factorization of v^μ and v that

$$\begin{aligned} \varphi_\mu(x, s, w) &= x + s - \left(\sqrt{\lambda_1(v^\mu)} u_1(v) + \sqrt{\lambda_2(v^\mu)} u_2(v)\right), \\ \varphi(x, s, w) &= x + s - \left(\sqrt{\lambda_1(v)} u_1(v) + \sqrt{\lambda_2(v)} u_2(v)\right), \end{aligned} \tag{40}$$

where

$$\begin{aligned} \lambda_i(v) = \|x\|^2 + \|s\|^2 + x^T s + w_0 + (-1)^i \|2x_0 x_1 + 2s_0 s_1 \\ + x_0 s_1 + s_0 x_1 + w_1\|, \end{aligned} \tag{41}$$

and $\lambda_i(v^\mu)$ and $u_i(v)$ are, respectively, given by (25) and (26) for $i = 1, 2$. It is obvious that

$$\lambda_i(v^\mu) = \lambda_i(v) + \mu^2, \tag{42}$$

for $i = 1, 2$. Then,

$$\begin{aligned} &\cdot \lim_{\mu \longrightarrow 0}\left(\sqrt{\lambda_1(v^\mu)} u_1(v) + \sqrt{\lambda_2(v^\mu)} u_2(v)\right) \\ &= \lim_{\mu \longrightarrow 0}\left(\sqrt{\lambda_1(v) + \mu^2} u_1(v) + \sqrt{\lambda_2(v) + \mu^2} u_2(v)\right) \\ &= \sqrt{\lambda_1(v)} u_1(v) + \sqrt{\lambda_2(v)} u_2(v), \end{aligned} \tag{43}$$

and $\lim_{\mu \longrightarrow 0} \varphi_\mu(x, s, w) = \varphi(x, s, w)$. Thus, by (i) and Definition 2, φ_μ is a smoothing function of φ.

(iii) By following the proof of Proposition 5.1 [15], we obtain the desired result. □

Next, we study some properties of φ, which will be used in the subsequent analysis.

Lemma 3. *For any $x = (x_0; x_1), s = (s_0; s_1)$, $\widetilde{w} = (\widetilde{w}_0; \widetilde{w}_1) \in \mathbb{R} \times \mathbb{R}^{n-1}$, let $x^2 + s^2 + \widetilde{w}^2 \in bd\mathcal{K}^n$. Then, we have*

$$\begin{aligned} x_0^2 &= \|x_1\|^2, \\ s_0^2 &= \|s_1\|^2, \\ \widetilde{w}_0^2 &= \|\widetilde{w}_1\|^2, \\ x_0 s_0 &= x_1^T s_1, \\ x_0 \widetilde{w}_0 &= x_1^T \widetilde{w}_1, \\ s_0 \widetilde{w}_0 &= s_1^T \widetilde{w}_1, \\ x_0 s_1 &= s_0 x_1, \\ x_0 \widetilde{w}_1 &= \widetilde{w}_0 x_1, \\ s_0 \widetilde{w}_1 &= \widetilde{w}_0^T s_1. \end{aligned} \tag{44}$$

Proof. We can obtain the desired result by following the proof of Lemma 2 [16]. □

Lemma 4. *For any $x = (x_0; x_1), s = (s_0; s_1) \in \mathbb{R} \times \mathbb{R}^{n-1}$, let $v = (v_0; v_1) = x^2 + s^2 + x \circ s + w \in bd\mathcal{K}^n$. Then, one has*

$$\begin{aligned} x \circ x' &= 0, \\ s \circ s' &= 0, \end{aligned} \tag{45}$$

$$\begin{aligned} x \circ s' &= 0, \\ x \circ \widetilde{w}\prime &= 0, \end{aligned} \tag{46}$$

$$\begin{aligned} s \circ \widetilde{w}\prime &= 0, \\ \widetilde{w} \circ \widetilde{w}\prime &= 0, \end{aligned} \tag{47}$$

$$\begin{aligned} x \circ v\prime &= 0, \\ s \circ v\prime &= 0, \end{aligned} \tag{48}$$

$$\begin{aligned} x_0^2 + s_0^2 + x_0 s_0 + \frac{w_0}{2} &= \left\| x_0 x_1 + s_0 s_1 + x_0 s_1 + \frac{w_1}{2} \right\| \\ &= \|x_1\|^2 + \|s_1\|^2 + x_1^T s_1 + \frac{w_1}{2}, \end{aligned} \tag{49}$$

where $\widetilde{w} := \sqrt{w}$. Moreover, the following equivalence holds:

$$\begin{aligned} &v_0 = 0 \Leftrightarrow v_1 = 0 \Leftrightarrow v = 0 \\ &\Leftrightarrow x_0 = s_0 = w_0 = 0 \Leftrightarrow x_1 = s_1 = w_1 = 0 \Leftrightarrow (x, s, w) = (0, 0, 0). \end{aligned} \tag{50}$$

Proof. Since

$$\begin{aligned} v &= x^2 + s^2 + x \circ s + w \\ &= \left(x + \frac{s}{2}\right)^2 + \frac{3}{4} s^2 + \widetilde{w}^2 \\ &= \left(s + \frac{x}{2}\right)^2 + \frac{3}{4} x^2 + \widetilde{w}^2 \in \mathrm{bd}\mathcal{K}^n, \end{aligned} \tag{51}$$

from Lemma 3, we have

$$\begin{aligned} \left(x + \frac{s}{2}\right) \circ \left(x + \frac{s}{2}\right)' &= 0, \quad s \circ s' = 0, \\ \left(x + \frac{s}{2}\right) \circ s' &= 0, \quad s \circ \widetilde{w}' = 0, \\ \left(x + \frac{s}{2}\right) \circ \widetilde{w}' &= 0, \quad \widetilde{w} \circ \widetilde{w}' = 0, \\ \left(s + \frac{x}{2} x\right) \circ \left(s + \frac{x}{2}\right)' &= 0, \quad x \circ x' = 0, \\ \left(s + \frac{x}{2}\right) \circ x' &= 0, \quad x \circ \widetilde{w}' = 0, \\ \left(s + \frac{x}{2}\right) \circ \widetilde{w}' &= 0. \end{aligned} \tag{52}$$

It follows from these equalities that the results in (45)–(47) hold. Since $v \in \mathrm{bd}\mathcal{K}^n$, we have $\lambda_1(v) = 0$, i.e.,

$$\|x\|^2 + \|s\|^2 + x^T s + w_0 = \|2x_0 x_1 + 2s_0 s_1 + x_0 s_1 + s_0 x_1 + w_1\|. \tag{53}$$

By the last relation and (45)–(47), we obtain that (49) holds. To prove (48), we only need to verify $x_0 v_1 = v_0 x_1$ and $x_1^T v_1 = x_0 v_0$ by the symmetry of x and s in v. From (45)–(47) and (49),

$$\begin{aligned} x_0 v_1 &= x_0 \left(2x_0 x_1 + 2s_0 s_1 + x_0 s_1 + s_0 x_1 + 2\widetilde{w}_0 \widetilde{w}_1\right) \\ &= 2\left(x_0^2 + s_0^2 + x_0 s_0 + \widetilde{w}_0^2\right) x_1 \\ &= 2\left(\|x_1\|^2 + \|s_1\|^2 + x_1^T s_1 + \frac{w_0}{2}\right) x_1 \\ &= v_0 x_1, \\ x_1^T v_1 &= x_1^T \left(2x_0 x_1 + 2s_0 s_1 + x_0 s_1 + s_0 x_1 + 2\widetilde{w}_0 \widetilde{w}_1\right) \\ &= 2x_0 \left(\|x_1\|^2 + s_0^2 + x_1^T s_1 + \widetilde{w}_0^2\right) \\ &= 2x_0 \left(x_0^2 + s_0^2 + x_0 s_0 + \frac{w_0}{2}\right) \\ &= x_0 v_0. \end{aligned} \tag{54}$$

From (51), the equivalence is also true. □

4. Jacobian Consistency

In this section, we will show the Jacobian consistency property and estimate the distance between the gradient of the smoothing function (7) and the subgradient of the WSOCCP complementarity function (5). For any $\mu \in \mathbb{R}$, $w \in \mathcal{K}^n$, let $z := (x, s, y) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m$. Based on smoothing function (7), we define Φ_μ: $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^{2n+m}$ by

$$\Phi(z) := \begin{pmatrix} F(x,s,y) \\ \varphi(x,s,w) \end{pmatrix}, \tag{55}$$

$$\Phi_\mu(z) := \begin{pmatrix} F(x,s,y) \\ \varphi_\mu(x,s,w) \end{pmatrix}. \tag{56}$$

From (1) and (56) and Lemma 1,

$$\Phi_\mu(z) = 0 \Leftrightarrow z = (x,s,y) \text{ solves WSOCCP (1).} \tag{57}$$

Since the function $\Phi(z)$ is typically nonsmooth, Newton's method cannot be applied to the system $\Phi(z) = 0$ directly. Thus, we can approximately solve the smooth system $\Phi_\mu(z) = 0$ at each iteration and make $\|\Phi_\mu(z)\|$ decrease gradually by reducing μ to zero. First, we show that the function $\Phi_\mu(z)$ satisfies the Jacobian consistency.

Lemma 5. *For any arbitrary but fixed vector $w \in \mathcal{K}^n$, we have for any $(\mu, x, s) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$,*

$$J_\varphi^0(x,s) := \lim_{\mu \longrightarrow 0} \varphi_\mu'(x,s,w) = \begin{pmatrix} I - L\left(x + \frac{s}{2}\right)J \\ \\ I - L\left(s + \frac{x}{2}\right)J \end{pmatrix}, \tag{58}$$

where

$$J := \begin{cases} L^{-1}(\omega), \text{ if } (x,s) \in \mathcal{I}, \\ \\ \frac{1}{2\sqrt{2v_0}}\left(1 \;\; \overline{v}_1^T \overline{v}_1 I - 3\overline{v}_1 \overline{v}_1^T\right), \text{ if } (x,s) \in \mathcal{B}, O, \text{ if } (x,s) \in \mathcal{O}. \end{cases} \tag{59}$$

Proof. By (34) and the symmetry of x and s, it suffices to prove

$$\lim_{\mu \longrightarrow 0} L\left(x + \frac{s}{2}\right) L^{-1}(\omega^\mu) = L\left(x + \frac{s}{2}\right) J. \tag{60}$$

□

Case 1. If $(x, s) \in \mathcal{I}$, it follows from (25) that

$$\begin{aligned} \lim_{\mu \longrightarrow 0} \omega^\mu &= \lim_{\mu \longrightarrow 0}\left[\sqrt{\lambda_1(v^\mu)}\, u_1(v) + \sqrt{\lambda_2(v^\mu)}\, u_2(v)\right] \\ &= \lim_{\mu \longrightarrow 0}\left[\sqrt{\lambda_1(v) + \mu^2}\, u_1(v) + \sqrt{\lambda_2(v) + \mu^2}\, u_2(v)\right] \\ &= \sqrt{\lambda_1(v)} u_1(v) + \sqrt{\lambda_2(v)} u_2(v) \\ &= \omega \in \operatorname{int}\mathcal{K}^n. \end{aligned} \tag{61}$$

Therefore,

$$\lim_{\mu \longrightarrow 0} L\left(x + \frac{s}{2}\right) L^{-1}(\omega^\mu) = L\left(x + \frac{s}{2}\right) L^{-1}(\omega). \tag{62}$$

Case 2. If $(x, s) \in \mathcal{B}$, it is easy to prove (51), and

$$\begin{aligned} 2v_0 &= \lambda_2(v) > \lambda_1(v) = 0, \\ \|v_1\| &= v_0 = \left\|x + \frac{s}{2}\right\|^2 + \frac{3}{4}\|s\|^2 + w_0 > 0. \end{aligned} \tag{63}$$

Thus, we obtain the following from (25):

$$\lambda_1(v^\mu) = \lambda_1(v) + \mu^2 = \mu^2 > 0, \tag{64}$$

$$\lambda_2(v^\mu) = \lambda_2(v) + \mu^2 = 2v_0 + \mu^2 > 0. \tag{65}$$

For any $\mu \neq 0$, we may get from (35) that $L^{-1}(\omega^\mu) = L_1(v^\mu) + L_2(v^\mu)$. We first prove for any $\mu \neq 0$,

$$L\left(x + \frac{s}{2}\right) L_1(v^\mu) = O. \tag{66}$$

Let

$$\vartheta := (1; \overline{v}_1) = \frac{1}{\|v_1\|}(v_0; v_1) = \frac{v}{v_0}. \tag{67}$$

Based on (36), (48), and (64), we have

$$\begin{aligned} L\left(x + \frac{s}{2}\right) L_1(v^\mu) &= \frac{1}{2\sqrt{\lambda_1(v^\mu)}} L\left(x + \frac{s}{2}\right) \vartheta \prime \vartheta'^T \\ &= \frac{1}{2|\mu|}\left(x + \frac{s}{2}\right) \circ \vartheta \prime \vartheta'^{\mathrm{T}} \\ &= \frac{1}{2|\mu| v_0^2}\left(x + \frac{s}{2}\right) \circ v \prime v'^{\mathrm{T}} \\ &= O. \end{aligned} \tag{68}$$

Next, we prove $\lim_{\mu \longrightarrow 0} L_2(v^\mu) = J$. From (37), (64), and (65), we have

$$\lim_{\mu\longrightarrow 0} L_2(v^\mu) = \lim_{\mu\longrightarrow 0} \frac{1}{2\sqrt{2v_0+\mu^2}}\begin{pmatrix} 1 & \overline{v}_1^T \\ \overline{v}_1 & \overline{v}_1\overline{v}_1^T \end{pmatrix} + \lim_{\mu\longrightarrow 0} \frac{2}{\sqrt{\mu^2}+\sqrt{2v_0+\mu^2}}\begin{pmatrix} 0 & 0^T \\ 0 & I-\overline{v}_1\overline{v}_1^T \end{pmatrix} = \frac{1}{2\sqrt{2v_0}}\begin{pmatrix} 1 & \overline{v}_1^T \\ \overline{v}_1 & 4I-3\overline{v}_1\overline{v}_1^T \end{pmatrix} = J. \tag{69}$$

Combining (68) and (69) yields

$$\lim_{\mu\longrightarrow 0} L\left(x+\frac{s}{2}\right)L^{-1}(\omega^\mu) = \lim_{\mu\longrightarrow 0} L\left(x+\frac{s}{2}\right)L_2(v^\mu) = L\left(x+\frac{s}{2}\right)J. \tag{70}$$

Case 3. If $(x,s)\in\mathcal{O}$, it follows from Lemma 4 that $(x,s,w)=(0,0,0)$ and

$$\omega^\mu = \sqrt{v^\mu} = |\mu|e \in \operatorname{int}\mathcal{K}^n,$$
$$\lim_{\mu\longrightarrow 0} L\left(x+\frac{s}{2}\right)L^{-1}(\omega^\mu) = \lim_{\mu\longrightarrow 0} O\cdot\frac{1}{|\mu|}e = O = L\left(x+\frac{s}{2}\right)J. \tag{71}$$

Lemma 6. *For any arbitrary but fixed vector $w\in\mathcal{K}^n$, we have for any $(x,s)\in\mathbb{R}^n\times\mathbb{R}^n$,*

$$\begin{pmatrix} I-U_x \\ I-U_s \end{pmatrix} \in \partial_B\varphi(x,s,w), \tag{72}$$

where

$$U_x = \pm\frac{1}{2}Z + L\left(x+\frac{s}{2}\right)J,$$
$$U_s = \pm Z + L\left(s+\frac{x}{2}\right)J,$$
$$Z = \begin{cases} O, & \text{if } (x,s)\in\mathcal{I}, \\ \frac{1}{2}\left(1 \quad -\overline{v}_1^T - \overline{v}_1\overline{v}_1^T\right), & \text{if } (x,s)\in\mathcal{B}, \\ I, & \text{if } (x,s)\in\mathcal{O}, \end{cases} \tag{73}$$

and J is defined by (59).

Proof. By Proposition 5.2 [15] and the chain rule for differentiation, the complementarity function φ is continuously differentiable at any $(x,s)\in\mathcal{I}$ with

$$\varphi\prime(x,s,w) = \begin{pmatrix} I-L\left(x+\frac{s}{2}\right)L^{-1}(\omega) \\ I-L\left(s+\frac{x}{2}\right)L^{-1}(\omega) \end{pmatrix} \in \partial_B\varphi(x,s,w). \tag{74}$$

Thus, it suffices to consider the two cases: $(x,s)\in\mathcal{B}$ and $(x,s)\in\mathcal{O}$.

For any $(x,s)\in\mathcal{B}$ or $(x,s)\in\mathcal{O}$, let $(x,\widehat{s}) = (x, s+\mu e)$ with sufficiently small $\mu\neq 0$, and define

$$\begin{aligned} \widehat{v} &= (\widehat{v}_0;\widehat{v}_1) := x^2+\widehat{s}^2+x\circ\widehat{s}+w, \\ \widehat{\omega} &= (\widehat{\omega}_0;\widehat{\omega}_1) := \sqrt{\widehat{v}}, \\ \widehat{\vartheta}_1 &:= \frac{\widehat{v}_1}{\|\widehat{v}_1\|}, \\ \widehat{\lambda}_i &= \lambda_i(\widehat{v}) := \widehat{v}_0 + (-1)^i\|\widehat{v}_1\|, \quad i=1,2. \end{aligned} \tag{75}$$

Then, we have

$$\begin{aligned} \widehat{v} &= x^2+(s+\mu e)^2+x\circ(s+\mu e)+w \\ &= v+\mu x+2\mu s+\mu^2 e, \\ \widehat{v}_0 &= v_0+\mu x_0+2\mu s_0+\mu^2, \\ \widehat{v}_1 &= v_1+\mu x_1+2\mu s_1, \end{aligned} \tag{76}$$

$$\widehat{\lambda}_i = v_0+\mu x_0+2\mu s_0+\mu^2+(-1)^i\|v_1+\mu x_1+2\mu s_1\|, \quad i=1,2. \tag{77}$$

Obviously, when $\mu\longrightarrow 0$, we have $(x,\widehat{s})\longrightarrow(x,s)$, $\widehat{v}\longrightarrow v$, $\widehat{\omega}\longrightarrow\omega$ and $\widehat{\lambda}_i\longrightarrow\lambda_i(v)$ for $i=1,2$. Then by (7), it suffices to show

$$\begin{aligned} \lim_{\mu\longrightarrow 0} L\left(x+\frac{\widehat{s}}{2}\right)L^{-1}(\widehat{\omega}) &= U_x, \\ \lim_{\mu\longrightarrow 0} L\left(\widehat{s}+\frac{x}{2}\right)L^{-1}(\widehat{\omega}) &= U_s, \end{aligned} \tag{78}$$

if φ is differentiable at $(x,\widehat{s})$. □

Case 4. If $(x,s)\in\mathcal{B}$, we obtain $v\in(\operatorname{bd}\mathcal{K}^n/\{0\})$, and from (45), (46), and (48),

$$\begin{aligned} \|\widehat{v}_1\|^2 &= \|v_1+\mu x_1+2\mu s_1\|^2 \\ &= \|v_1\|^2+\mu^2\|x_1\|^2+4\mu^2\|s_1\|^2+4\mu v_1^T s_1 \\ &\quad + 2\mu v_1^T x_1 + 4\mu^2 x_1^T s_1 \\ &= (v_0+\mu x_0+2\mu s_0)^2. \end{aligned} \tag{79}$$

The last relation together with $v_0>0$ implies that for sufficiently small μ, we have

$$\|\widehat{v}_1\| = v_0 + \mu x_0 + 2\mu s_0 > 0. \tag{80}$$

For sufficiently small $\mu \neq 0$, we obtain from (77) and (80),

$$\widehat{\lambda}_1 = v_0 + \mu x_0 + 2\mu s_0 + \mu^2 - \|\widehat{v}_1\| = \mu^2 > 0, \tag{81}$$

$$\widehat{\lambda}_2 = v_0 + \mu x_0 + 2\mu s_0 + \mu^2 + \|\widehat{v}_1\| = 2(v_0 + \mu x_0 + 2\mu s_0) + \mu^2 > 0. \tag{82}$$

It follows from (81) and (82) that $\widehat{v} \in \operatorname{int}\mathscr{K}^n$, and hence φ is differentiable at $(x, \widehat{s})$.

Now we will prove

$$\lim_{\mu \longrightarrow 0} L\left(x + \frac{\widehat{s}}{2}\right) L^{-1}(\widehat{\omega}) = U_x, \tag{83}$$

where $L^{-1}(\widehat{\omega}) = L_1(\widehat{v}) + L_2(\widehat{v})$, in which $L_1(\widehat{v})$ and $L_2(\widehat{v})$ are given by (36) and (37) with $\widehat{v}$ and $\widehat{\vartheta}_1$ replacing v^μ and $\overline{v}_1$, respectively. By the expression of $\widehat{v}_1$ and (80),

$$\begin{aligned}
\widehat{\vartheta} &:= \left(1; \widehat{\vartheta}_1\right) = \frac{1}{\|\widehat{v}_1\|}\left(\|\widehat{v}_1\|; \widehat{v}_1\right) \\
&= \frac{1}{\|\widehat{v}_1\|}\left(v_0 + \mu x_0 + 2\mu s_0; v_1 + \mu x_1 + 2\mu s_1\right) \\
&= \frac{1}{\|\widehat{v}_1\|}(v + \mu x + 2\mu s).
\end{aligned} \tag{84}$$

By (45), (46), (48), and (84), we have

$$\begin{aligned}
\left(x + \frac{s}{2}\right) \circ \widehat{\vartheta}\prime &= \frac{1}{\|\widehat{v}_1\|}\left(x + \frac{s}{2}\right) \circ (v + \mu \mathrm{x} + 2\mu \mathrm{s})\prime \\
&= \frac{1}{\|\widehat{v}_1\|}\left[\left(x + \frac{s}{2}\right) \circ v\prime + 2\mu\left(\mathrm{x} + \frac{\mathrm{s}}{2}\right) \circ \left(\frac{\mathrm{x}}{2} + \mathrm{s}\right)\prime\right] \\
&= 0.
\end{aligned} \tag{85}$$

Thus, from (36) and (81),

$$\begin{aligned}
L\left(x + \frac{\widehat{s}}{2}\right) L_1(\widehat{v}) &= \frac{1}{2\sqrt{\widehat{\lambda}_1}}\left(x + \frac{s}{2} + \frac{\mu e}{2}\right) \circ \widehat{\vartheta}\prime \widehat{\vartheta}'^{\mathrm{T}} \\
&= \frac{1}{2|\mu|}\left[\left(x + \frac{s}{2}\right) \circ \widehat{\vartheta}\prime \widehat{\vartheta}'^{\mathrm{T}} + \frac{\mu}{2}\widehat{\vartheta}\prime \widehat{\vartheta}'^{\mathrm{T}}\right] \\
&= \frac{\operatorname{sgn}(\mu)}{4}\widehat{\vartheta}\prime \widehat{\vartheta}'^{T}.
\end{aligned} \tag{86}$$

It follows from (73)–(84) that as $\mu \longrightarrow 0$,

$$\begin{aligned}
&\widehat{\lambda}_1 \longrightarrow \lambda_1(v) = 0, \\
&\widehat{\lambda}_2 \longrightarrow \lambda_2(v) = 2v_0, \\
&\widehat{\vartheta}_1 \longrightarrow \overline{v}_1, \\
&\frac{1}{2}\widehat{\vartheta}\prime \widehat{\vartheta}'^{T} = \frac{1}{2}\begin{pmatrix} 1 & -\widehat{\vartheta}_1^T \\ -\widehat{\vartheta}_1 & - - \hat{\vartheta}_1 \widehat{\vartheta}_1^T \end{pmatrix} \longrightarrow Z.
\end{aligned} \tag{87}$$

Then, by following the proof of Case 5 in Lemma 5, we have

$$\lim_{\mu \longrightarrow 0} L_2(\widehat{v}) = \frac{1}{2\sqrt{2v_0}}\begin{pmatrix} 1 & \overline{v}_1^T \\ \overline{v}_1 & 4I - 3\overline{v}_1\overline{v}_1^T \end{pmatrix} = J. \tag{88}$$

Therefore, we obtain from (86) and (88) that

$$\begin{aligned}
\lim_{\mu \longrightarrow \pm 0} L\left(x + \frac{\widehat{s}}{2}\right) L^{-1}(\widehat{\omega}) &= \lim_{\mu \longrightarrow \pm 0} L\left(x + \frac{\widehat{s}}{2}\right) L_1(\widehat{v}) \\
&\quad + \lim_{\mu \longrightarrow \pm 0} L\left(x + \frac{\widehat{s}}{2}\right) L_2(\widehat{v}) \\
&= \lim_{\mu \longrightarrow \pm 0} \frac{\operatorname{sgn}(\mu)}{4}\widehat{\vartheta}\prime \widehat{\vartheta}'^{T} + L\left(x + \frac{s}{2}\right) J \\
&= \pm \frac{1}{2} Z + L\left(x + \frac{s}{2}\right) J \\
&= U_x.
\end{aligned} \tag{89}$$

Next we will prove

$$\lim_{\mu \longrightarrow 0} L\left(\widehat{s} + \frac{x}{2}\right) L^{-1}(\widehat{\omega}) = U_s. \tag{90}$$

By (45), (46), (48), (81), and (84), we have

$$\begin{aligned}
\left(s + \frac{x}{2}\right) \circ \widehat{\vartheta}\prime &= \frac{1}{\|\widehat{v}_1\|}\left(s + \frac{x}{2}\right) \circ (v + \mu \mathrm{x} + 2\mu \mathrm{s})\prime \\
&= \frac{1}{\|\widehat{v}_1\|}\left[\left(s + \frac{x}{2}\right) \circ v\prime + 2\mu\left(\mathrm{s} + \frac{\mathrm{x}}{2}\right) \circ \left(\frac{\mathrm{x}}{2} + \mathrm{s}\right)\prime\right] \\
&= 0,
\end{aligned} \tag{91}$$

and then

$$L\left(\hat{s}+\frac{x}{2}\right)L_1(\hat{v}) = \frac{1}{2\sqrt{\hat{\lambda}_1}}\left(s+\mu e+\frac{x}{2}\right)\circ\hat{\vartheta}_I\hat{\vartheta}'^{\mathrm{T}}$$
$$= \frac{1}{2|\mu|}\left[\left(s+\frac{x}{2}\right)\circ\hat{\vartheta}_I\hat{\vartheta}'^{\mathrm{T}} + \mu\hat{\vartheta}_I\hat{\vartheta}'^{\mathrm{T}}\right] \tag{92}$$
$$= \frac{\operatorname{sgn}(\mu)}{2}\hat{\vartheta}_I\hat{\vartheta}'^{T}.$$

Therefore, we obtain from (88) and (92) that

$$\lim_{\mu\longrightarrow\pm 0} L\left(\hat{s}+\frac{x}{2}\right)L^{-1}(\hat{\omega}) = \lim_{\mu\longrightarrow\pm 0} L\left(\hat{s}+\frac{x}{2}\right)L_1(\hat{v})$$
$$+ \lim_{\mu\longrightarrow\pm 0} L\left(\hat{s}+\frac{x}{2}\right)L_2(\hat{v})$$
$$= \lim_{\mu\longrightarrow\pm 0} \frac{\operatorname{sgn}(\mu)}{2}\hat{\vartheta}_I\hat{\vartheta}'^{T} + L\left(s+\frac{x}{2}\right)J$$
$$= \pm Z + L\left(s+\frac{x}{2}\right)J$$
$$= U_s. \tag{93}$$

Case 5. If $(x,s)\in\mathcal{O}$, it follows from Lemma 4 that $(x,s,w) = (0,0,0)$. Thus, $\hat{v} = \mu^2 e \in \operatorname{int}\mathcal{K}^n$, $\hat{\omega} = |\mu|e$, and

$$\lim_{\mu\longrightarrow\pm 0} L\left(x+\frac{\hat{s}}{2}\right)L^{-1}(\hat{\omega}) = \lim_{\mu\longrightarrow\pm 0}\frac{\mu}{2}I\cdot\frac{1}{|\mu|}I = \lim_{\mu\longrightarrow\pm 0}\frac{\operatorname{sgn}(\mu)}{2}I$$
$$= \pm\frac{1}{2}I = U_x,$$
$$\lim_{\mu\longrightarrow\pm 0} L\left(\hat{s}+\frac{x}{2}\right)L^{-1}(\hat{\omega}) = \lim_{\mu\longrightarrow\pm 0}\mu I\cdot\frac{1}{|\mu|}I = \lim_{\mu\longrightarrow\pm 0}\operatorname{sgn}(\mu)I$$
$$= \pm I = U_s. \tag{94}$$

□

Now we show the Jacobian consistency of the function Φ_μ (56) and then estimate an upper bound of the parameter $\mu > 0$ for the predicted accuracy of the distance between the gradient of Φ_μ (56) and the subgradient of Φ (55).

Theorem 1. *The following results hold. (i) The function Φ_μ defined by (56) with $\mu > 0$ satisfies the Jacobian consistency. (ii) For given $\tau > 0$ and any point $z := (x,s,y)\in\mathbb{R}^{2n+m}$, let $\rho(x,s)$ be any function such that*

$$\rho(x,s) \geq \left\| \begin{matrix} L\left(x+\frac{s}{2}\right)J \\ L\left(s+\frac{x}{2}\right)J \end{matrix} \right\|, \tag{95}$$

and let $\overline{\mu}$: $\mathbb{R}^{2n}\times\mathbb{R}_+ \longrightarrow \mathbb{R}_+\cup\{+\infty\}$ be defined by

$$\overline{\mu}(x,s,\tau) := \begin{cases} \dfrac{\lambda_1(v)\tau}{\sqrt{\rho^2(x,s)-\lambda_1(v)\tau^2}}, & \text{if } (x,s)\in\mathcal{I} \text{ and } \tau < \left(\rho(x,s)/\sqrt{\lambda_1(v)}\right), \\[2ex] \dfrac{v_0\tau}{\sqrt{2\rho(x,s)\left(2\rho(x,s)-\tau\sqrt{2v_0}\right)}}, & \text{if } (x,s)\in\mathcal{B} \text{ and } \tau < 2\rho(x,s)/\sqrt{2v_0}, \\[2ex] +\infty, & \text{otherwise.} \end{cases} \tag{96}$$

Then, for any $\mu\in\mathbb{R}$ such that $0<|\mu|\leq\overline{\mu}(x,s,\tau)$, we have

$$\operatorname{dist}\left(\Phi_{\mu_I}(z), \partial\Phi(z)\right) < \tau. \tag{97}$$

Proof. By (56), it suffices to show the Jacobian consistency of φ_μ with $\mu > 0$. Define

$$V^i := \begin{pmatrix} I - U_x^i \\ I - U_s^i \end{pmatrix}, \tag{98}$$

where

$$U_x^i = (-1)^i\frac{1}{2}Z + L\left(x+\frac{s}{2}\right)J,$$
$$U_s^i = (-1)^i Z + L\left(s+\frac{x}{2}\right)J, \tag{99}$$

for $i = 1,2$, J and Z are defined by (59) and (73). Let

$$V := \frac{1}{2}\left(V^1 + V^2\right) = \begin{pmatrix} I - L\left(x + \frac{s}{2}\right)J \\ \\ I - L\left(s + \frac{x}{2}\right)J \end{pmatrix}. \tag{100}$$

It follows from Lemma 5 and Lemma 6 that

$$V = J_\varphi^0(x, s) = \lim_{\mu \longrightarrow 0} \varphi_{\mu'}(x, s, w), \tag{101}$$

and $V^1, V^2 \in \partial_B \varphi(x, s, w)$. Hence,

$$V = \frac{1}{2}\left(V^1 + V^2\right) \in \partial\varphi(x, s, w), \tag{102}$$

which together with Definition 1 and Lemma 2 implies the Jacobian consistency of φ_μ with $\mu > 0$. (ii) For any $z := (x, s, y) \in \mathbb{R}^{2n+m}$, it follows from the proof of Theorem 1(i) that

$$\begin{aligned} &J_\varphi^0(x, s) = V \in \partial\varphi(x, s, w), \\ &J_\Phi^0(z) := \begin{pmatrix} J_\varphi^0(x, s) & O \\ F_{x,s}'(x, s, y) & F_y'(x, s, y) \end{pmatrix} \in \partial\Phi(x, s, y). \end{aligned} \tag{103}$$

Thus, we obtain from (34) and (100) that

$$\begin{aligned} \operatorname{dist}\left(\Phi_\mu'(z), \partial\Phi(z)\right) &\le \left\|\Phi_\mu'(z) - J_\Phi^0(z)\right\| \\ &= \left\|\varphi_\mu'(z) - J_\varphi^0(z)\right\| \\ &= \left\| \begin{matrix} L\left(x + \frac{s}{2}\right)\left(L^{-1}(\widehat{\omega}) - J\right) \\ \\ L\left(s + \frac{x}{2}\right)\left(L^{-1}(\widehat{\omega}) - J\right) \end{matrix} \right\|. \end{aligned} \tag{104}$$

Then, similar to the proof of Proposition 4.1 [13], we have

$$\operatorname{dist}\left(\Phi_\mu'(z), \partial\Phi(z)\right) \le \left|g_0(x, s) - g_\mu(x, s)\right| \cdot \left\| \begin{matrix} L\left(x + \frac{s}{2}\right)J \\ \\ L\left(s + \frac{x}{2}\right)J \end{matrix} \right\|, \tag{105}$$

where g_μ: $\mathbb{R}^{2n} \longrightarrow \mathbb{R}_+$ is given by

$$g_\mu(x, s) := \begin{cases} \dfrac{1}{\sqrt{\lambda_1(v) + \mu^2}}, & \text{if } (x, s) \in \mathscr{I}, \\ \\ \dfrac{2}{\sqrt{2v_0 + \mu^2} + |\mu|} \dfrac{1}{\sqrt{\lambda_1(v) + \mu^2}}, & \text{if } (x, s) \in \mathscr{B}, \\ \\ 0, & \text{if } (x, s) \in \mathscr{O}. \end{cases} \tag{106}$$

Hence, by following the proof of Theorem 4.1 [13], the result holds. □

5. Conclusions

In this paper, we show the Jacobian consistency of the smoothing function φ_μ for WSOCCP, which will play a key role in analyzing the rapid convergence of smoothing methods. Moreover, in order to adjust a parameter appropriately in smoothing methods, we estimate the distance between the gradient of the smoothing function φ_μ and the subgradient of the weighted SOC complementarity function φ.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (no. 11861026), Guangxi Key Laboratory of Cryptography and Information Security (no. GCIS201819), and Guangxi Key Laboratory of Automatic Detecting Technology and Instruments, China (no. YQ18112).

References

[1] F. A. Potra, "Weighted complementarity problems---a new paradigm for computing equilibria," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1634–1654, 2012.

[2] K. Anstreicher, "Interior-point algorithms for a generalization of linear programming and weighted centring," *Optimization Methods and Software*, vol. 27, no. 4-5, pp. 605–612, 2012.

[3] J. Tang, "A variant nonmonotone smoothing algorithm with improved numerical results for large-scale LWCPs," *Computational and Applied Mathematics*, vol. 37, no. 3, pp. 3927–3936, 2018.

[4] X. Chi, M. S. Gowda, and J. Tao, "The weighted horizontal linear complementarity problem on a Euclidean Jordan algebra," *Journal of Global Optimization*, vol. 73, no. 1, pp. 153–169, 2019.

[5] S. Pan and J.-S. Chen, "A semismooth Newton method for SOCCPs based on a one-parametric class of SOC complementarity functions," *Computational Optimization and Applications*, vol. 45, no. 1, pp. 59–88, 2010.

[6] J.-S. Chen and S. Pan, "A one-parametric class of merit functions for the second-order cone complementarity problem," *Computational Optimization and Applications*, vol. 45, no. 3, pp. 581–606, 2010.

[7] L. Qi, D. Sun, and G. Zhou, "A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities," *Mathematical Programming*, vol. 87, no. 1, pp. 1–35, 2000.

[8] S. Hayashi, N. Yamashita, and M. Fukushima, "A combined smoothing and regularization method for monotone second-

order cone complementarity problems," *SIAM Journal on Optimization*, vol. 15, no. 2, pp. 593–615, 2005.

[9] N. Krejić, S. Rapajić, Globally convergent jacobian smoothing inexact Newton methods for NCP," *Computational Optimization and Applications*, vol. 41, no. 2, pp. 243–261, 2008.

[10] P. Chen, P. Zhang, X. Zhu X et al., "Modified Jacobian smoothing method for nonsmooth complementarity problems," *Computational Optimization and Applications*, vol. 75, no. 1, pp. 205–305, 2020.

[11] J. Faraut and A. Korányi, *Analysis on Symmetric Cones*, Oxford University Press, New York, NY, USA, 1994.

[12] X. Chen, L. Qi, and D. Sun, "Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities," *Mathematics of Computation*, vol. 67, no. 222, pp. 519–541, 1998.

[13] H. Ogasawara and Y. Narushima, "The Jacobian consistency of a smoothed Fischer-Burmeister function associated with second-order cones," *Journal of Mathematical Analysis and Applications*, vol. 394, no. 1, pp. 231–247, 2012.

[14] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, NY, USA, 1983.

[15] M. Fukushima, Z. Luo, and P. Tseng, "Smoothing functions for second-order-cone complementarity problems," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 436–460, 2001.

[16] J. Chen and P. Tseng, "An unconstrained smooth minimization reformulation of the second-order cone complementarity problem," *Mathematical ProgrammingD*, vol. 104, no. 2-3, pp. 293–327, 2005.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6614177, 7 pages
https://doi.org/10.1155/2020/6614177

Research Article

A Class of Optimal Liquidation Problem with a Nonlinear Temporary Market Impact

Jiangming Ma [1] **and Di Gao** [2]

[1] *School of Economics, Xihua University, Chengdu, Sichuan 610039, China*
[2] *School of International Business, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China*

Correspondence should be addressed to Di Gao; gaodi1593@163.com

Received 14 October 2020; Revised 29 November 2020; Accepted 11 December 2020; Published 24 December 2020

Academic Editor: Mohamed El Ghami

We extend the self-exciting model by assuming that the temporary market impact is nonlinear and the coefficient of the temporary market impact is an exponential function. Through optimal control method, the optimal strategy satisfies the second-order nonlinear ordinary differential equation. The specific form of the optimal strategy is given, and the decreasing property of the optimal strategy is proved. A numerical example is given to illustrate the financial implications of the model parameter changes. We find that the optimal strategy of a risk-neutral investor changes with time and investment environment.

1. Introduction

In the financial field, the problem of optimal liquidation is widely studied. In 1998, Bertsimas and Lo [1] study the minimum transaction completion in the case of fixed trading time dynamic trading strategy. Based on the original scholar's model, Almgren and Chriss [2] consider the expected costs and risks of execution and propose a simple market impact model. It includes the following three parts: unaffect price process, temporary market impact, and permanent. Almgren–Chriss market impact model provides a good tool to continue studying the optimal liquidation problem. Almgren [3] gives the optimal execution strategy under the nonlinear temporary impact. Curato et al. [4] study the optimal execution of a large trade when the transient impact is nonlinear. Gueant and Lehalle [5] carry out research on optimal liquidation when the execution process intensity is general functional forms.

Some scholars research the corresponding optimal liquidation strategy under the expanded Almgren–Chriss model. Schied and Gatheral [6, 7] show the optimal strategies when the unaffected price process is geometric Brownian motion. Lehalle and Neuman [8] obtain the optimal strategies and provide the existence and uniqueness of them when the model incorporates a Markovian signal. When order flow is imbalanced and uncertain, the optimal execution is discussed by Bechler and Ludkovski [9] and Cheng et al. [10], respectively. Cartea and Jaimungal [11] and Gueant et al. [12] address the optimal liquidation when the order book is limited. Many scholars continue to make further research studies in the recent years. Cartea and Jaimungal [13] investigate optimal execution when the investor executes a large order. Kato [14] gets the optimal execution of trader when the volume weighted average price (VWAP) is used in the Almgren–Chriss model. Frei and Westray [15] propose a relative volume curve model under the VWAP model and get the explicit characterization of optimal execution. Based on [14], Kato [16] obtains a second-order asymptotic expansion formula of optimal strategies by the penalization method. Klöck et al. [17] change the application scenario and study the execution with dark pool in the Almgren–Chriss model. Bela et al. [18] study the optimal liquidation under the Almgren–Chriss model with running and terminal inventory costs and general predictive signals about price changes. Bank et al. [19] carry out research on the optimal problem of hedging and give the general predictable target hedging strategies. In addition, some scholars investigate the optimal liquidation by using

new methods. Damian [20] discusses the optimal execution under the multitime version of the Almgren–Chriss model by the variational calculus techniques which assumes that the optimal control is in the set of admissible controls. Bismuth et al. [21] address the optimal liquidation in an Almgren–Chriss framework by the Bayesian learning and dynamic programming techniques when expected returns are unknown. Besides, Schied and Zhang [22] consider the Almgren–Chriss model has n risk-averse agents and prove the property of optimal liquidation strategies.

Differential equations are widely used in engineering. Wakif et al. [23, 24] study the stability of nanofluids which has the characteristics of electrically conducting and Newtonian fluids, incorporating the effects of thermophoresis and Brownian motion in different situations. Then, they get the corresponding differential equations which are obtained by the relevant methods. From the numerical methods, they discuss the properties and get the solutions of differential equations. Similarly, differential equations are also used in the financial field. The optimal strategies of relevant literature mentioned above satisfy the differential equations through the optimal control methods.

Caye and Muhle-Karbe [25] consider that the trades not only incur price impact but also increase the execution costs. Thus, they propose a self-exciting price model and get the optimal liquidation strategies under the Almgren–Chriss framework. However, they only discuss the temporary impact, and its coefficients are linear functions. Different from the above references, we suppose that the temporary impact and its coefficients are nonlinear functions. Namely, let the temporary impact and its coefficient be the exponential function and the power function, respectively, which are used in economic and finance research. Finally, we get the specific form and prove the properties of optimal liquidation.

The paper is organized as follows. In Section 2, we state the Almgren–Chriss framework, self-exciting price model, and objective function. In Section 3, we give the specific form of optimal liquidation and discuss the properties of solutions. In Section 4, we show the numerical examples and the corresponding financial interpretations.

2. Statement of Background

In this paper, we use the continuous-time market impact model of Almgren–Chriss which supposes that the active time of every investor is fixed in $[0,T]$. An investor hold x shares at the initial time and completely trade at the time T, that is, $X_0 = x$ and $X_T = 0$. The investor's strategy is X_t which is absolutely continuous and bounded with derivative $\dot{X}_t$ and $X_t = x + \int_0^t \dot{X}_t \mathrm{d}t$, where $\dot{X}_t$ satisfies $\int_0^T (\dot{X}_t)^2 < \infty$.

A filtration $(\mathcal{F}_t)_{t>0}$ on the given probability space $(\Omega, \mathcal{F}, \mathcal{P})$ is supported by the standard Brownian motion W_t. We suppose that the a risk asset's unaffected price process follows the Bachelier [26] model with no drift:

$$S_t^0 := \sigma\, dW_t. \tag{1}$$

The Almgren–Chriss model is supposed that the price of a risk asset is related to the hold share and trading speed at the time t. So, the Almgren–Chriss market impact model is divided into three components: unaffected price process, permanent impact components, and temporary impact components. The specific form of the Almgren–Chriss model is assumed to be

$$S_t := S_t^0 + \gamma(X_t - x) + \lambda \dot{X}_t, \tag{2}$$

where $\gamma(X_t - x)$ and $\dot{X}_t$ represents the permanent impact and temporary impact components, respectively; the parameters $\lambda > 0$ and $\gamma > 0$ represent the coefficient of permanent and temporary impact components.

Cayé and Muhle-Karbe [25] give the self-exciting price impact under the Almgren–Chriss framework. In this model, the parameter of temporary impact component is a linear function about the number of shares already sold. The specific form is assumed to be

$$S_t := S_t^0 + (a + b(x - X_t))\dot{X}_t, \quad t \in [0,T], \tag{3}$$

where $a > 0$ and $b > 0$. In equation (3), there is no permanent impact component because the influence of the permanent impact component about the cost of investor is fixed.

At each time $t \in [0,T]$, the infinitesimal amounts of $-\dot{X}_t\, \mathrm{d}t$ shares are sold at price S_t. Therefore, the total implementation cost is represented by

$$C(X) := xS_0 + \int_0^T \dot{X}_t\, S_t \mathrm{d}t. \tag{4}$$

So, the optimal trade execution problem becomes the minimization of expected costs. We only need to solve the minimization of expected cost:

$$\text{minimize}\, \mathbb{E}[C(X)]. \tag{5}$$

Problem (5) is proposed by Bertsimas and Lo [1]. Carmona and Yang [27] use (5) to deal with the problem of the maximization.

3. Main Results

Cayé and Muhle-Karbe [25] only discuss the coefficient of temporary impact component is a linear function. However, in the real lifetime, the coefficient of temporary impact component maybe nonlinear. So, we suppose that the coefficient of temporary impact component like the exponential function is widely used in economic activities. Thus, the coefficient of temporary impact component is assumed to be $e^{a+b(x-X_t)}$, where $a > 0$ and $b > 0$.

Theorem 1. *Since the coefficient of temporary impact component is $e^{a+b(x-X_t)}$, there exists a unique strategy for mean optimization. The strategy is the unique solution of the following differential equation:*

$$\ddot{X}_t - \frac{b}{2}\dot{X}_t^2 = 0, \tag{6}$$

with two-point boundary conditions

$$X_0 = x,$$
$$X_T = 0. \tag{7}$$

The solution of equation (6) is

$$t = C_1 + C_2 \int e^{-(b/2)X_t} \mathrm{d}X_t. \tag{8}$$

Proof. When the coefficient of temporary impact component is $e^{a+b(x-X_t)}$, equation (3) is

$$S_t := S_t^0 + e^{a+b(x-X_t)} \dot{X}_t, \quad t \in [0, T]. \tag{9}$$

From equations (4) and (9), we obtain

$$\begin{aligned} C(X) &:= xS_0 + \int_0^T \dot{X}_t S_t \mathrm{d}t \\ &= xS_0 + \int_0^T \dot{X}_t \left(S_t^0 + \left(e^{a+b(x-X_t)} \dot{X}_t \right) \mathrm{d}t \right) \\ &= \int_0^T \sigma X_t dW_t + \int_0^T e^{a+b(x-X_t)} \dot{X}_t^2 \mathrm{d}t. \end{aligned} \tag{10}$$

From the properties of Ito integral and equation (5), we obtain

$$\text{minimize}\, \mathbb{E}[C(X)] = \int_0^T e^{a+b(x-X_t)} \dot{X}_t^2 \mathrm{d}t. \tag{11}$$

In order to get the solution of equation (11), we use the Euler–Lagrange equation to get the second-order ordinary differential equation:

$$\ddot{X}_t - \frac{b}{2} \dot{X}_t^2 = 0. \tag{12}$$

The optimal strategy satisfies equation (12). From [28], the solution of equation (12) is

$$t = C_1 + C_2 \int e^{-(b/2)X_t} \mathrm{d}X_t. \tag{13}$$

□

Theorem 2. *The optimal strategy from equations (12) and (11), over all deterministic, absolutely continuous strategies X_t, is decreasing.*

Proof

$$C(X) = \int_0^T e^{a+b(x-X_t)} \dot{X}_t^{\,2} \mathrm{d}t = e^{a+bx} \int_0^T e^{-X_t} \dot{X}_t^{\,2} \mathrm{d}t = \mathrm{F}(X). \tag{14}$$

Let $Y = X - X^*$; then, we obtain

$$\begin{aligned} \mathscr{F}(X) &= \mathscr{F}(Y + X^*) \\ &= e^{a+bx} \int_0^T e^{-X_t^* - Y_t} \left(\dot{X}_t^* + \dot{Y}_t \right)^2 \mathrm{d}t \\ &= e^{a+bx} \int_0^T e^{-X_t^* - Y_t} \left(\dot{X}_t^{*\,2} + 2\dot{X}_t^* \dot{Y}_t + \dot{Y}_t^{\,2} \right) \mathrm{d}t \\ &= e^{a+bx} \int_0^T e^{-X_t^* - Y_t} \dot{X}_t^{*\,2} \mathrm{d}t + e^{a+bx} \int_0^T e^{-X_t^* - Y_t} 2\dot{X}_t^* \dot{Y}_t \, \mathrm{d}t \\ &\quad + e^{a+bx} \int_0^T e^{-X_t^* - Y_t} \dot{Y}_t^{\,2} \mathrm{d}t \\ &\geq e^{a+bx} \int_0^T e^{-X_t^* - Y_t} \dot{X}_t^{*\,2} \mathrm{d}t \\ &\geq e^{a+bx} \int_0^T e^{-X_t^*} \dot{X}_t^{*\,2} \mathrm{d}t. \end{aligned} \tag{15}$$

Because the Almgren–Chriss model supposes that there is no existence price manipulation, the $\dot{X}_t$ satisfies $\dot{X}_t < 0$. Thus, X_t is decreasing. From the properties of exponential function and integral, we get the proof of Theorem 2.

Except for references [3, 5, 10], there are still some scholars studying the problem of optimal execution when temporary impact is nonlinear. Gatheral [29] discusses the optimal liquidation problems under the basic assumption of the Almgren–Chriss model which contains some special nonlinear temporary market impact function. When the temporary market impact function in the Almgren–Chriss model is nonlinear, Labadie and Lehalle [30] examine the optimal starting times, stopping times, and risk measures for algorithmic trading of target close and implementation shortfall. Hendricks and Wilcox [31] research the optimal trade execution of the Almgren–Chriss framework by a reinforcement learning method. Horst and Naujokat [32] show the value derivatives under market impact in a multiplayer framework which is based on the nonlinear temporary market impact function of the Almgren–Chriss model.

Although Caye and Muhle-Karbe [25] pay attention to the optimal liquidation of self-exciting price impact under the Almgren–Chriss framework, the case of nonlinear temporary market impact function is not studied. Next, we suppose that the temporary market impact function likes a nonlinear form $h(\dot{X}_t)$. Since $h(\dot{X}_t)$ has many forms, drawing on the above research studies of optimal liquidation with the nonlinear functions, we let $h(\dot{X}_t)$ be the power function. Namely, $h(\dot{X}_t)$ has the form $h(\dot{X}_t) = (\dot{X}_t)^\alpha$, $\alpha > 0$. However, in the actual process of

solution, it is difficult to get the general solution of optimizing equation when the $h(\dot{X}_t)$ is the power function. Therefore, we research the special case which is usually used in economic and finance and discuss the optimal strategies when $h(\dot{X}_t) = (\dot{X}_t)^2$. Thus, equation (3) is changed for

$$S_t := S_t^0 + (a + b(x - X_t))(\dot{X}_t)^2, \quad t \in [0, T]. \tag{16}$$

□

Theorem 3. *Since the temporary impact component is $h(\dot{X}_t) = (\dot{X}_t)^2$, there exists a unique strategy for mean optimization. The strategy is the unique solution of the following differential equation:*

$$3(a + b(x - X_t))\ddot{X}_t - b\dot{X}_t^2 = 0, \tag{17}$$

with two-point boundary conditions

$$\begin{aligned} X_0 &= x, \\ X_T &= 0. \end{aligned} \tag{18}$$

The solution of equation (17) is

$$(a + b(x - X_t))^{(2/3)} = C_1 t + C_2. \tag{19}$$

Proof. When the coefficient of temporary impact component is $e^{a+b(x-X_t)}$, equation (3) is

$$S_t := S_t^0 + (a + b(x - X_t))(\dot{X}_t)^2, \quad t \in [0, T]. \tag{20}$$

From equations (4) and (20), we obtain

$$\begin{aligned} C(X) &:= xS_0 + \int_0^T \dot{X}_t S_t \mathrm{d}t \\ &= xS_0 + \int_0^T \dot{X}_t \left(S_t^0 + (a + b(x - X_t))(\dot{X}_t)^2\right)\mathrm{d}t \\ &= \int_0^T \sigma X_t \mathrm{d}W_t + \int_0^T (a + b(x - X_t))\dot{X}_t^3 \mathrm{d}t. \end{aligned} \tag{21}$$

By the properties of Ito integral and equation (5), we obtain

$$\text{minimize}\, \mathbb{E}[C(X)] = \int_0^T (a + b(x - X_t))\dot{X}_t^3 \mathrm{d}t. \tag{22}$$

In order to get the solution of equation (22), we use the Euler–Lagrange equation to get the second-order ordinary differential equation. The optimal strategy satisfies the following equation:

$$3(a + b(x - X_t))\ddot{X}_t - b\dot{X}_t^2 = 0. \tag{23}$$

From [28], the solution of equation (23) is

$$(a + b(x - X_t))^{(2/3)} = C_1 t + C_2. \tag{24}$$

□

Theorem 4. *The optimal strategy from equations (17) and (24), over all deterministic, absolutely continuous strategies X_t, is decreasing.*

Proof

$$\begin{aligned} C(X) &= \int_0^T (a + b(x - X_t))\dot{X}_t^3 \\ &\quad \mathrm{d}t = (a + bx)\int_0^T (-bX_t)\dot{X}_t^3 \mathrm{d}t = \mathscr{F}(X). \end{aligned} \tag{25}$$

Let $Y = X - X^*$; then, we obtain

$$\begin{aligned} \mathscr{F}(X) = \mathscr{F}(Y + X^*) &= (a + bx)\int_0^T (-bX_t^* - bY_t)\left(\dot{X}_t^* + \dot{Y}_t\right)^3 \mathrm{d}t \\ &= (a + bx)\int_0^T (-bX_t^* - bY_t)\left(\dot{X}_t^{*3} + 3\dot{X}_t^{*2}\dot{Y}_t \right. \\ &\quad \left. + 3\dot{X}_t^*\dot{Y}_t^2 + \dot{Y}_t^3\right)\mathrm{d}t \\ &= (a + bx)\int_0^T (-bX_t^* - bY_t)\dot{X}_t^{*2}\mathrm{d}t \\ &\quad + (a + bx)\int_0^T (-bX_t^* - bY_t)\left(3\dot{X}_t^{*2}\dot{Y}_t\right)\mathrm{d}t \\ &\quad + (a + bx)\int_0^T (-bX_t^* - bY_t)\left(3\dot{X}_t^*\dot{Y}_t^2\right)\mathrm{d}t \\ &\quad + (a + bx)\int_0^T (-bX_t^* - bY_t)\left(\dot{Y}_t^3\right)\mathrm{d}t \\ &\geq (a + bx)\int_0^T (-bX_t^* - bY_t)\dot{X}_t^{*3}\mathrm{d}t \\ &\geq (a + bx)\int_0^T \left(-bX_t^*\dot{X}_t^{*3}\right)\mathrm{d}t. \end{aligned} \tag{26}$$

From the properties of $\dot{X}_t$, $\dot{X}_t < 0$, and integral, we get the proof of Theorem 4.

Next, we discuss the optimal liquidation strategies when the temporary impact function is power function and the coefficient of temporary impact is $e^{a+b(x-X_t)}$. Thus, the price process is changed to be

$$S_t := S_t^0 + e^{a+b(x-X_t)}(\dot{X}_t)^2, \quad t \in [0, T]. \tag{27}$$

□

Theorem 5. *Since the temporary impact component is $h(\dot{X}_t) = (\dot{X}_t)^2$ and the coefficient of temporary impact component is $e^{a+b(x-X_t)}$, there exists a unique strategy for mean optimization. The strategy is the unique solution of the following differential equation:*

$$3\ddot{X}_t - b\dot{X}_t^2 = 0, \tag{28}$$

with two-point boundary conditions

$$\begin{aligned} X_0 &= x, \\ X_T &= 0. \end{aligned} \tag{29}$$

The solution of equation (6) is

$$(a + b(x - X_t))^{(2/3)} = C_1 t + C_2. \tag{30}$$

Proof. From equations (4) and (19), we obtain

$$\begin{aligned} C(X) &:= xS_0 + \int_0^T \dot{X}_t\, S_t \mathrm{d}t, \\ &= xS_0 + \int_0^T e^{a+b(x-X_t)} \left(\dot{X}_t\right)^3 \mathrm{d}t \\ &= \int_0^T \sigma X_t \mathrm{d}W_t + \int_0^T e^{a+b(x-X_t)} \dot{X}_t^{\ 3} \mathrm{d}t. \end{aligned} \tag{31}$$

Through the properties of Ito integral and equation (5), we have

$$\text{minimize}\, \mathbb{E}[C(X)] = \int_0^T e^{a+b(x-X_t)} \dot{X}_t^{\ 3} \mathrm{d}t. \tag{32}$$

In order to get the solution of equation (32), we use the Euler–Lagrange equation to get the second-order ordinary differential equation:

$$3\ddot{X}_t - b\dot{X}_t^{\ 2} = 0. \tag{33}$$

The optimal strategy satisfies equation (33). From [28], the solution of equation (33) is

$$t = C_1 + C_2 \int e^{-(b/3)X_t} \mathrm{d}X_t. \tag{34}$$

□

Theorem 6. *The optimal strategy from equations (28) and (34), over all deterministic, decreasing, and absolutely continuous strategies x with square-integrable derivative, satisfies $X_0 = x$ and $X_T = 0$.*

Proof

$$\begin{aligned} C(X) &= \int_0^T e^{a+b(x-X_t)} \dot{X}_t^{\ 3} \mathrm{d}t \\ &= (a+bx) \int_0^T (-bX_t) \dot{X}_t^{\ 3} \mathrm{d}t = \mathcal{F}(X). \end{aligned} \tag{35}$$

Let $Y = X - X^*$; then, we obtain

$$\begin{aligned} \mathcal{F}(X) = \mathcal{F}(Y + X^*) &= e^{(a+bx)} \int_0^T e^{(-bX_t^* - bY_t)} \left(\dot{X}_t^* + \dot{Y}_t\right)^3 \mathrm{d}t \\ &= e^{(a+bx)} \int_0^T e^{(-bX_t^* - bY_t)} \left(\dot{X}_t^{*\ 3} + 3\dot{X}_t^{*\ 2} \dot{Y}_t \right. \\ &\quad \left. + 3\dot{X}_t^* \dot{Y}_t^{\ 2} + \dot{Y}_t^{\ 3}\right) \mathrm{d}t \\ &= e^{(a+bx)} \int_0^T e^{(-bX_t^* - bY_t)} \dot{X}_t^{*\ 2} \mathrm{d}t \\ &\quad + (a+bx) \int_0^T e^{(-bX_t^* - bY_t)} \left(3\dot{X}_t^{*\ 2} \dot{Y}_t\right) \mathrm{d}t \\ &\quad + e^{(a+bx)} \int_0^T e^{(-bX_t^* - bY_t)} \left(3\dot{X}_t^* \dot{Y}_t^{\ 2}\right) \mathrm{d}t \\ &\quad + e^{(a+bx)} \int_0^T e^{(-bX_t^* - bY_t)} \left(\dot{Y}_t^{\ 3}\right) \mathrm{d}t \\ &\geq e^{(a+bx)} \int_0^T e^{(-bX_t^* - bY_t)} \dot{X}_t^{*\ 3} \mathrm{d}t \\ &\geq e^{(a+bx)} \int_0^T e^{-bX_t^*} \dot{X}_t^{*\ 3} \mathrm{d}t. \end{aligned} \tag{36}$$

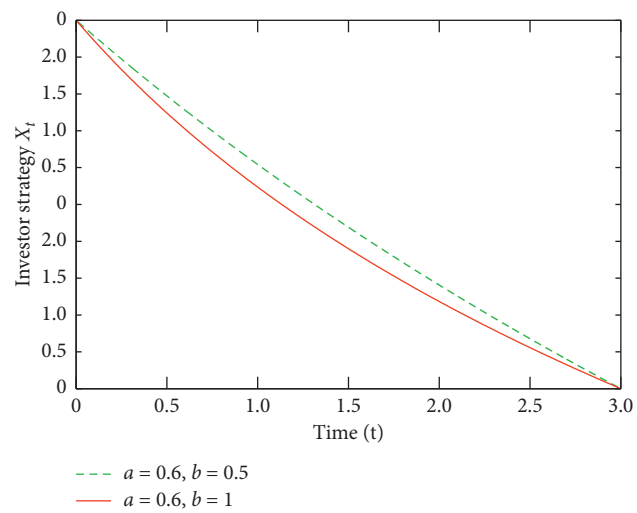


Figure 1: Equation (13).

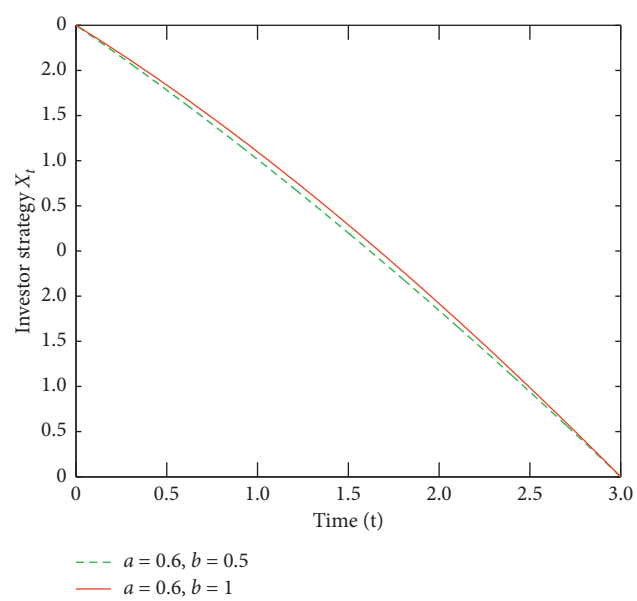


Figure 2: Equation (24).

From the properties of $\dot{X}_t$, $\dot{X}_t < 0$, and integral, we get the proof of Theorem 6. □

4. Numerical Simulation

In the previous part, we give the specific forms of optimal investment strategies for risk-neutral investors when the temporary market impact and coefficient of temporary market impact are a power function and an exponential function, respectively. According to the parameter setting method in the relevant literature, we assume that $X_0 = 2$, $X_{T=3} = 0$, and $t \in [0, 3]$, and the values of other parameters are shown in the figures.

From equation (13), we know that the optimal liquidation has nothing to do with a. From Figure 1, we get that the cost of trading becomes higher when b gets larger. Thus,

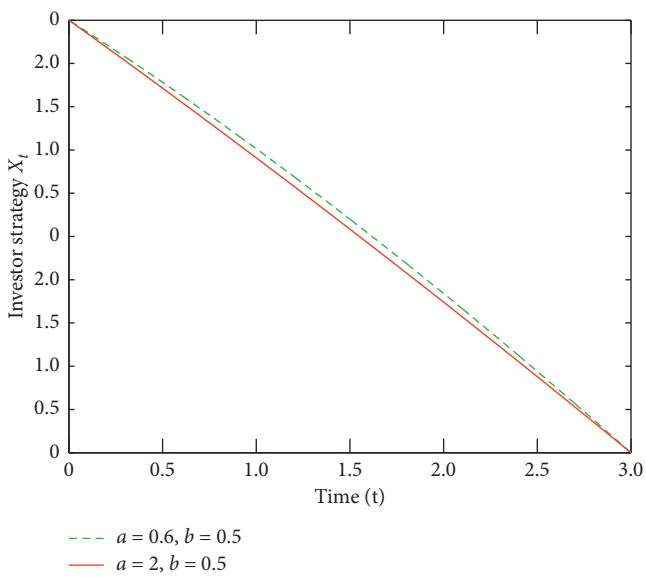


Figure 3: Equation (24).

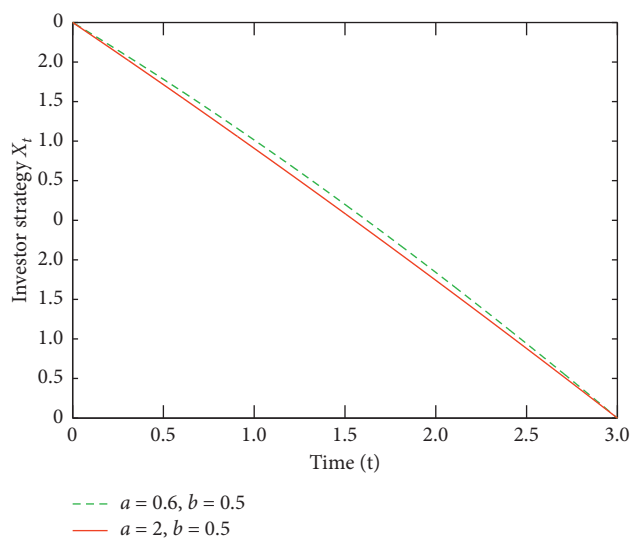


Figure 4: Equation (34).

the investor speeds up liquidation early. From equation (24) and Figures 2 and 3, when the temporary market impact is a power function and the coefficient of temporary market impact is a linear function, the investor realizes that they will face large execution costs with bigger a and smaller b so that they speed up liquidation early and slow down the trading speed later. When temporary market impact is a power function and the coefficient of temporary market impact is an exponential function, we find that a has no effect on the optimal liquidation. In Figure 4, the bigger b leads to increased costs of execution. Therefore, in order to decrease the costs, the investor will speed up liquidation early.

Through numerical examples, we find that when investment conditions change, the optimal investment strategy of risk-neutral investors is not the average of initial holdings with respect to time. However, it changes with time and investment environment. Since investment environment is complicated and volatile, the purpose of this paper is to remind risk-neutral investors that when they face the three investment environments, and they should follow these investment strategies to get the maximum return.

5. Conclusion

In this paper, combining the model setting of Cay and Muhle-Karbe [25] with the review of relevant literature, we put forward a class of optimal liquidation when the temporary market impact is a power function and the coefficient of temporary market impact is an exponential function, respectively. The optimal liquidation strategies satisfy the second-order nonlinear ordinary differential equations. The form of optimal liquidation strategies is given. At the same time, we discuss the properties of optimal liquidation strategies. Through the numerical example, we explain the financial implications with the changed parameter. This paper studies the optimal liquidation strategy of investors under three situations. In the future, more situations with financial implications will be discussed, particularly the fractional form of derivation with financial implications.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant 61903064), Key Research Base of Philosophy and Social Sciences for Colleges and Universities in Sichuan Province (KJJR2019-004), and Talent Introduction Project of Xihua University (w202247).

References

[1] D. Bertsmas and A. Lo, "Optimal execution of execution costs," *Journal of Finance Markets*, vol. 1, pp. 1–50, 1998.
[2] R. Almgren and N. Chriss, "Optimal execution of portfolio transactions," *Journal of Risk*, vol. 3, no. 2, pp. 5–39, 2000.
[3] R. F. Almgren, "Optimal execution with nonlinear impact functions and trading-enhanced risk," *Applied Mathematical Finance*, vol. 10, no. 1, pp. 1–18, 2003.
[4] G. Curato, J. Gatheral, and F. Lillo, "Optimal execution with non-linear transient market impact," *Quantitative Finance*, vol. 17, no. 1, pp. 41–54, 2017.
[5] O. Gueant and C. A. Lehalle, "General intersity shapes in optimal liquidation," *Mathematical Finance*, vol. 25, no. 3, pp. 457–495, 2015.
[6] A. Schied, "Robust strategies for optimal order execution in the almgren-chriss framework," *Applied Mathematical Finance*, vol. 20, no. 3, pp. 264–286, 2013.

[7] J. Gatheral and A. Schied, "Optimal trade execution under geometric brownian motion in the almgren and Chriss framework," *International Journal of Theoretical and Applied Finance*, vol. 14, no. 3, pp. 353–368, 2011.
[8] C.-A. Lehalle and E. Neuman, "Incorporating signals into optimal trading," *Finance and Stochastics*, vol. 23, no. 2, pp. 275–311, 2019.
[9] K. Bechler and M. Ludkovski, "Optimal execution with dynamic order flow imbalance," *SIAM Journal on Financial Mathematics*, vol. 6, no. 1, pp. 1123–1151, 2015.
[10] X. Cheng, M. D. Giacinto, and T. Wang, "Optimal execution with uncertain order fills in Almgren-Chriss framework," *Quantitative Finance*, vol. 17, no. 1, pp. 55–69, 2017.
[11] A. Cartea and S. Jaimungal, "Optimal execution with limit and market orders," *Quantitative Finance*, vol. 15, no. 8, pp. 1279–1291, 2015.
[12] O. Gueant, C.-A. Lehalle, and J. Fernandez-Tapia, "Optimal portfolio liquidation with limit orders," *SIAM Journal on Financial Mathematics*, vol. 3, no. 1, pp. 740–764, 2012.
[13] A. Cartea and S. Jaimungal"Incorporating order-flow into optimal execution," *Mathematics and Financial Economics*, vol. 10, no. 3, pp. 339–364, 2016.
[14] T. Kato, "VWAP execution as an optimal strategy," *JSIAM Letters*, vol. 7, pp. 33–36, 2015.
[15] C. Frei and N. Westray, "Optimal Execution of A VWAP order: a stochastic control approach," *Mathematical Finance*, vol. 25, no. 3, pp. 612–639, 2015.
[16] T. Kato, "An optimal execution problem in the volume-dependent Almgren-Chriss model," *Algorithmic Finance*, vol. 17, pp. 1–14, 2018.
[17] F. Klöck, A. Schied, and Y. Sun, "Price manipulation in a market impact model with dark pool," *Applied Mathematical Finance*, vol. 24, no. 5, pp. 417–450, 2017.
[18] C. Bela, J. Muhle-Karbe, and K. Ou, "Liquidation in target zone models," *Market Microstructure and Liquidity*, vol. 4, no. 3, pp. 1–12, 2018.
[19] P. Bank, H. M. Soner, and M Vo, "Hedging with temporary price impact," *Mathematics and Financial Economics*, vol. 11, no. 2, pp. 215–239, 2017.
[20] V. Voß, "Modelling optimal execution strategies for Algorithmic trading," *Theoretical and Applied Economics*, vol. XXII, no. 4, pp. 99–104, 2015.
[21] A. Bismuth, O. Gueant, and J. Pu, "Portfolio choice, portfolio liquidation, and portfolio transition under drift uncertainty," *Mathematics and Financial Economics*, vol. 13, no. 4, pp. 661–719, 2019.
[22] A. Schied and T. Zhang, "A state-constrained differential game arising in optimal portfolio liquidation," *Mathematical Finance*, vol. 27, no. 3, pp. 779–802, 2017.
[23] A. Wakif, Z. Boulahia, F. Ali, M. R. Eid, and R. Sehaqui, "Numerical analysis of the unsteady natural convection MHD Couette nanofluid flow in the presence of thermal radiation using single and two-phase nanofluid models for CuCWater nanofluids," *International Journal of Applied and Computational Mathematics*, vol. 4, no. 81, pp. 1–27, 2018.
[24] A. Wakif, A. Chamkha, T. Thumma, I. L. Animasaun, and R. Sehaqui, "Thermal radiation and surface roughness efects on the thermo-magneto-hydrodynamic stability of alumina-copper oxide hybrid nanofuids utilizing the generalized Buongiorno's nanofuid model," *Journal of Thermal Analysis and Calorimetry*, 2020.
[25] T. Cayé and J. Muhle-Karbe, "Liquidation with self-exciting price impact," *Mathmatical Finance Economic*, vol. 10, no. 17, pp. 15–28, 2016.
[26] L. Bachelier, "Théorie de la spéculation," *Annales scientifiques de l'École normale supérieure*, vol. 17, no. 17, pp. 21–86, 1900.
[27] R. A. Carmona and Z. Joseph Yang, "Optimal liquidation under stochastic price impact," *International Journal of Theoretical and Applied Finance*, vol. 21, no. 8, pp. 1–28, 2018.
[28] D. P. Andrei and F. Z. Valentin, *Handbook of Exact Solutions for Ordinary Differential Equations*, Chapman Hall/CRC, Boca Raton, FL, USA, 2003.
[29] J. Gatheral, "No-dynamic-arbitrage and market impact," *Quantitative Finance*, vol. 10, no. 7, pp. 749–759, 2010.
[30] M. Labadie and C. Lehalle, "Optimal starting times, stopping times and risk measures for algorithmic trading: target close and implementation shortfall," 2013, https://arxiv.org/pdf/1205.3482.pdf.
[31] D. Hendricks and D. Wilcox, "A reinforcement learning extension to thealmgren-chriss framework for optimal trade execution," 2014, https://arxiv.org/pdf/1403.2229.pdf.
[32] U. Horst and F. Naujokat, "Illiquidity and derivative valuation," 2018, https://arxiv.org/pdf/0901.0091.pdf.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6642725, 13 pages
https://doi.org/10.1155/2020/6642725

Research Article

A Double Nonmonotone Quasi-Newton Method for Nonlinear Complementarity Problem Based on Piecewise NCP Functions

Zhensheng Yu,[1] **Zilun Wang**,[1] **and Ke Su**[2]

[1]*College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China*
[2]*College of Mathematics and Information Science, Hebei University, Hebei 071002, China*

Correspondence should be addressed to Zhensheng Yu; zhsh-yu@163.com

Received 22 October 2020; Revised 12 November 2020; Accepted 26 November 2020; Published 16 December 2020

Academic Editor: Guoqiang Wang

In this paper, a double nonmonotone quasi-Newton method is proposed for the nonlinear complementarity problem. By using 3-1 piecewise and 4-1 piecewise nonlinear complementarity functions, the nonlinear complementarity problem is reformulated into a smooth equation. By a double nonmonotone line search, a smooth Broyden-like algorithm is proposed, where a single solution of a smooth equation at each iteration is required with the reduction in the scale of the calculation. Under suitable conditions, the global convergence of the algorithm is proved, and numerical results with some practical applications are given to show the efficiency of the algorithm.

1. Introduction

In this paper, we consider the following nonlinear complementarity problem (NCP): find $x \in R^n$ such that

$$x \geq 0, F(x) \geq 0, x^T F(x) = 0. \tag{1}$$

where F: $R^n \longrightarrow R^n$ is continuously differentiable and the superscript T denotes the transpose operator. When F is linear, problem (1) reduces to a linear complementarity problem (LCP). Throughout this paper, the solution set of problem (1), denoted by X^*, is assumed to be nonempty.

Nonlinear complementarity problems arisen in many practical applications, for example, the KKT systems of mathematical programming problem, the economic equilibrium, the engineering design problem, can be reformulated into the NCP [1–3].

During the past decades, various efficient numerical algorithms are proposed to solve the NCP. One of the most effective methods is to transform the NCP into the semismooth equations (based on nonlinear complementarity function, NCP function) so that the semismooth Newton-type method can be deployed. The most well-known NCP functions are the Fischer–Burmeister function [4] (FB NCP function) and the modified FB NCP function [5]. Sun and Qi [6] proposed several NCP functions, investigated their properties, and provided a numerical comparison between the behavior of different NCP functions. Based on NCP functions, some kinds of algorithm are designed, see, for example, [7–11].

Another well-known class of algorithm is the smoothing algorithm. The main idea of smoothing algorithm is to reformulate the NCP to smooth equations by introducing the smoothing NCP functions. Some smoothing NCP functions and the corresponding algorithms can be found in [12–15].

Besides the NCP functions mentioned above, a 3-1 piecewise NCP function was proposed by Liu et al. [16], using it to solve the inequality-constrained nonlinear optimization. The advantage of the 3-1 piecewise lies in the absence of the smoothing parameter. Motivated by the 3-1 piecewise NCP function, Su and Yang [17, 18] developed smooth-based Newton algorithms with nonmonotone line search for nonlinear complementarity and generalized nonlinear complementarity problems. Different from the

previous methods, the authors introduced independent variable quantities to simplify the algorithm, reducing the amount of calculation without using the smoothing parameter.

Smoothing procedure allows one to use successful quasi-Newton approaches, and there are many quasi Newton methods available for the nonlinear complementarity problems based on some smoothing functions [19–26].

In this paper, we will construct a 3-1 piecewise and 4-1 piecewise NCP functions and develop a double nonmonotone quasi Newton method to solve the nonlinear complementarity problems. Based on the piecewise NCP functions, the nonlinear complementarity problem is transformed into the smooth equation. Moreover, we only solve one smooth equation at each iteration. In order to get the better numerical results, a double nonmonotone line search is used by combining with the Broyden-like algorithm. Consequently, the omission of the parameter μ and the single calculation of the Jacobian matrix at each iteration have led to the simplicity and flexibility of this approach. Furthermore, let $t = F(x)$ as an independent variable, which has no relationship with x, ensures the realization of our algorithm easier. Our algorithm is proved to be well-defined and globally convergent under suitable conditions. At the end of the paper, we give numerical results to prove the effectiveness of the algorithm. This paper is organized as follows: the piecewise linear NCP functions are introduced in Section 1. The double nonmonotone line search with quasi-Newton method is given in Section 2. In Section 3, the convergence properties of the algorithm are presented. We give some numerical results in Section 4, and the conclusion is drawn in Section 5.

2. Algorithm Analysis

To describe our algorithm, we first give the definitions of NCP function and P_0 function. We assume that $F\colon R^n \longrightarrow R^n$ is a continuously differentiable P_0-function; if, for all $x, y \in R^n$ with $x \neq y$, there exists an index i such that

$$(x_i - y_i)^T [F_i(x) - F_i(y)] \geq 0, \quad x_i \neq y_i, \tag{2}$$

and we regard a pair $(a, b) \in R^2$ as an NCP pair if $a \geq 0, b \geq 0$, and $a^T b = 0$; a function $\Phi\colon R^2 \longrightarrow R$ is called an NCP function, and we have $\Phi(a, b) = 0$ if and only if (a, b) is a NCP pair.

In what follows, we first introduce the 3–1 piecewise NCP function and then define a 4–1 piecewise NCP function:

$$\Phi(a,b) = \begin{cases} 3a - \left(\dfrac{a^2}{b}\right), & b \geq a > 0 \quad \text{or} \quad 3b > -a \geq 0; \\[2ex] 3a - \left(\dfrac{b^2}{a}\right), & a > b > 0 \quad \text{or} \quad 3a > -b \geq 0; \\[2ex] 9a + 9b, & \text{else.} \end{cases} \tag{3}$$

If $(a, b) \neq (0, 0)$, then

$$\nabla\Phi(a,b) = \begin{cases} \begin{pmatrix} 3 - \left(\dfrac{2a}{b}\right) \\[2ex] \left(\dfrac{a^2}{b^2}\right) \end{pmatrix}, & b \geq a > 0, \quad \text{or} \quad 3b > -a \geq 0; \\[2ex] \begin{pmatrix} \left(\dfrac{b^2}{a^2}\right) \\[2ex] 3 - \left(\dfrac{2b}{a}\right) \end{pmatrix}, & a > b > 0 \quad \text{or} \quad 3a > -b \geq 0; \\[2ex] \begin{pmatrix} 9 \\ 9 \end{pmatrix}, & \text{else.} \end{cases} \tag{4}$$

We define the 4–1 piecewise linear NCP function (k is any positive integer):

$$\Phi(a,b) = \begin{cases} k^2 a, & \text{if} \quad b \geq k|a|; \\[2ex] 2kb - \left(\dfrac{b^2}{a}\right), & \text{if} \quad a > \dfrac{|b|}{k}; \\[2ex] 2k^2 a + 2kb + \left(\dfrac{b^2}{a}\right), & \text{if} \quad a < -\dfrac{|b|}{k}; \\[2ex] k^2 a + 4kb, & \text{if} \quad b \leq -k|a| < 0. \end{cases} \tag{5}$$

If $(a, b) \neq (0, 0)$, then

$$\nabla\Phi(a,b)=\begin{cases}\begin{pmatrix}k^2\\0\end{pmatrix}, & \text{if}\quad b\geq k|a|;\\[2ex] \begin{pmatrix}\left(\dfrac{b^2}{a^2}\right)\\[2ex] 2k-\left(\dfrac{2b}{a}\right)\end{pmatrix}, & \text{if}\quad a>\dfrac{|b|}{k};\\[2ex] \begin{pmatrix}2k^2-\left(\dfrac{b^2}{a^2}\right)\\[2ex] 2k+\left(\dfrac{2b}{a}\right)\end{pmatrix}, & \text{if}\quad a<-\dfrac{|b|}{k};\\[2ex] \begin{pmatrix}k^2\\4k\end{pmatrix}, & \text{if}\quad b\leq -k|a|<0.\end{cases}$$

Denote $H\colon R^{2n}\longrightarrow R^{2n}$,

$$H(x,t)=\begin{pmatrix}t-F(x)\\ \Phi(x,t)\end{pmatrix}, \tag{6}$$

where t is a sequence in the algorithm and $t=F(x)$ holds at the optimal solution to NCP.

Hence, the NCP can be written as the following minimization problem:

$$\min\Psi(x,t)=\|H(x,t)\|. \tag{7}$$

To get the solution of (7), we introduce the notations as follows:

$$\left(\alpha_i^k,\beta_i^k\right)=\begin{cases}(1,1), & (x,t)=(0,0);\\ \nabla\Phi(x,t), & \text{otherwise.}\end{cases} \tag{8}$$

$i=1,2,\ldots,n.$Obviously, $\alpha_i^k>0$and$\beta_i^k>0$.

Denote the Jacobian matrix of $H(x^k,t^k)$ by $V(x^k,t^k)$, we get

$$V\left(x^k,t^k\right)=\begin{pmatrix}-F\prime\left(x^k\right) & I\\ \operatorname{diag}\left(\alpha_i^k\right) & \operatorname{diag}\left(\beta_i^k\right)\end{pmatrix}. \tag{9}$$

The identity matrix of $n\times n$, diagonal matrix whose ith diagonal element is α_i^k, and the diagonal matrix whose ith diagonal element is β_i^k are represented by I, $\operatorname{diag}(\alpha_i^k)$, and $\operatorname{diag}(\beta_i^k)$, respectively.

We use the nonmonotone line search to present Broyden-like method. The search directions d and λ are obtained by calculating a system of smooth equation, and the algorithm is described in detail in Algorithm1.

3. Convergence Analysis

In this section, the global convergence properties of a Broyden-like algorithm with 3–1 piecewise NCP function are discussed. We give some assumptions to prove the convergence of the algorithm.

Assumption 1

(a) Suppose $F\colon R^n\longrightarrow R^n$is P_0-function and it is continuously differentiable.

(b) On the level set of

$$L\left(x^0,t^0\right)=\left\{(x,t)\in R^{2n}|\Psi t(x,t)n\leq q\Psi h\left(x^0,t^0\right)\right\}, \tag{10}$$

where F is Lipschitz continuously differentiable, namely, there exists a constant Lsuch that for all $x_1,x_2\in R^n$,

$$\left\|F(x_1)-F(x_2)\right\|\leq L\left\|x_1-x_2\right\|. \tag{11}$$

Remark 1 (see [27]). $F(x)$ is P_0-function, then $F'(x)$ is positive semidefinite.

Lemma 1. *If $H(x^0,t^0)\neq 0$, then $B_0=V_0$ is nonsingular.*

Proof. Assume $H(x^0,t^0)\neq 0$. If $V_0^T(u,v)=0$ for some $(u,v)\in R^{2n}$, where $u=(u_1,u_2,\ldots,u_n)^T$ and $v=(v_1,v_2,\ldots,v_n)^T$, then

$$-F\prime\left(x^0\right)u+Iv=0, \tag{12}$$

$$\operatorname{diag}\left(\alpha^0\right)u+\operatorname{diag}\left(\beta^0\right)v=0. \tag{13}$$

By the definitions of α_i^0 and β_i^0, for all i, $\alpha_i^0>0$ and $\beta_i^0>0$. Therefore, $\operatorname{diag}(\beta^0)$ is nonsingular. Then

$$v=-\left(\operatorname{diag}\left(\beta^0\right)\right)^{-1}\operatorname{diag}\left(\alpha^0\right)u. \tag{14}$$

Substitute v in (12) by (14), and multiply by u^T, we have

$$-u^TF\left(x^0\right)u-u^T\left(\operatorname{diag}\left(\beta^0\right)\right)^{-1}\operatorname{diag}\left(\alpha^0\right)u=0. \tag{15}$$

According to the definition of P_0-function, all the principal minor determinants of $F\prime(x)$ is nonnegative; hence, $F\prime(x)$ is positive semidefinite. And matrix $(\operatorname{diag}(\beta^0))^{-1}\operatorname{diag}(\alpha^0)$ is positive definite. Therefore $u=0$. Together with (14), it holds that $v=0$, which implies B_0 is nonsingular. □

Lemma 2. *Assume that Assumption 1 holds. Then $\Phi(x^k,t^k)\longrightarrow 0$, as $k\longrightarrow\infty$.*

Proof. For convenience, we define $\|\Phi^{l(k)}\|=\max_{0\leq r\leq m(k)-1}\|\Phi^{k-r}\|$, where $k-m(k)+1\leq l(k)\leq k$. When $m(k+1)\leq m(k)+1$, we have

$$\begin{aligned}\left\|\Phi^{l(k+1)}\right\| &= \max_{0\le r\le m(k+1)-1}\left\|\Phi^{k+1-r}\right\| \\ &\le \max_{0\le r\le m(k)}\left\|\Phi^{k+1-r}\right\| \\ &= \max\left\{\left\|\Phi^{l(k)}\right\|, \left\|\Phi^{k+1}\right\|\right\} \\ &= \left\|\Phi^{l(k)}\right\|.\end{aligned} \tag{16}$$

Which means $\|\Phi^{l(k)}\|$ is decreasing monotonely, and hence, we have $\left\{\|\Phi^{l(k)}\|\right\}$ convergent. Based on (c) of Algorithm1, we have $\|\Phi^{l(k)}\| \le \|\Phi^{l(l(k)-1)}\|$.

By $\xi \in (0,1)$, $\left\{\|\Phi^{l(k)}\|\right\} \longrightarrow 0\,(k \longrightarrow \infty)$ holds, so according to $\|\Phi^{k+1}\| \le \xi\|\Phi^{l(k)}\| \longrightarrow 0$, the conclusion holds. □

Lemma 3. *Assume Assumption 1 holds. Then $t^k - F(x^k) \longrightarrow 0$ as $k \longrightarrow \infty$.*

Proof. Define $\|t^{l(k)} - F(x^{l(k)})\| = \max_{0\le r\le m(k)-1}\|t^{k-r} - F(x^{k-r})\|$, where $k - M \le l(k) \le k$. For $m(k+1) \le m(k) + 1$, we have

$$\begin{aligned}\left\|t^{l(k+1)} - F\left(x^{l(k+1)}\right)\right\| &= \max_{0\le r\le m(k+1)-1}\left\|t^{k+1-r} - F\left(x^{k+1-r}\right)\right\| \\ &\le \max_{0\le r\le m(k)}\left\|t^{k+1-r} - F\left(x^{k+1-r}\right)\right\| \\ &= \max\left\{\left\|t^{l(k)} - F\left(x^{l(k)}\right)\right\|, \left\|t^{k+1} - F\left(x^{k+1}\right)\right\|\right\} \\ &= \left\|t^{l(k)} - F\left(x^{l(k)}\right)\right\|.\end{aligned} \tag{17}$$

From (17), $\|t^{l(k)} - F(x^{l(k)})\|$ is decreasing in a monotone way; then $\{\|t^{l(k)} - F(x^{l(k)})\|\}$is convergent.

According to (g) of Algorithm 1, $\|t^{l(k)} - F(x^{l(k)})\| \le \xi\|t^{l(l(k)-10)} - F(x^{l(l(k)-1)})\|$. By $\xi \in (0,1)$, $\{\|t^{l(k)} - F(x^{l(k)})\|\} \longrightarrow 0\,(k \longrightarrow \infty)$ holds. That means $\|t^{k+1} - F(x^{k+1})\| \le \xi\|t^{l(k)} - F(x^{l(k)})\| \longrightarrow 0$ holds by Algorithm1, so the conclusion is as follows. □

Lemma 4. *Assume Assumption 1 holds. Then $d^k \longrightarrow 0$, $\lambda^k \longrightarrow 0$, and $H^k \longrightarrow 0$, as $k \longrightarrow \infty$.*

Proof. We have $\Phi(x^k, t^k) \longrightarrow 0$, $[t^k - F(x^k)] \longrightarrow 0$, as $k \longrightarrow \infty$ by Lemma 2 and Lemma 3.

So, $H(x^k, t^k) \longrightarrow 0$, as $k \longrightarrow \infty$:

$$B_k\begin{pmatrix} d^k \\ \lambda^k \end{pmatrix} = \begin{pmatrix} F\left(x^k\right) - t^k \\ -\Phi\left(x^k, t^k\right) \end{pmatrix} = 0. \tag{18}$$

We know that B_k is nonsingular by Algorithm1. So, $d^k \longrightarrow 0$, and $\lambda^k \longrightarrow 0$, as $k \longrightarrow \infty$. □

Theorem 1. *Under the same condition in Lemma 4, equation (a) of Algorithm1 has solutions, and the definition of Algorithm1 is well.*

Proof. On the one hand, we know B_0 is nonsingular by Lemma 1. And B_k produced by the Broyden-like iteration is nonsingular. Hence equation (a) of Algorithm1 has one and only one solution. On the other hand, we know $\Phi(x^k, t^k) \longrightarrow 0$ and $[t^k - F(x^k)] \longrightarrow 0$ as $k \longrightarrow \infty$ by Lemma 2 and Lemma 3. So $\|H(x^k, t^k)\| \longrightarrow 0$ as $k \longrightarrow \infty$. □

Lemma 5. *Assume Assumption 1 holds, and let $\{(x^k, t^k)\}$ be generated sequence by Algorithm1; then $\{(x^k, t^k)\} \subset L(x^0, t^0)$.*

Proof. By induction, for $k = 0$, we have $(x^0, t^0) \in L(x^0, t^0)$.

Assume $(x^k, t^k) \in L(x^0, t^0)$; then we have $\Psi(x^k, t^k) \le \Psi(x^0, t^0)$. By (c) and (d) of Algorithm1, we get

$$\begin{aligned}\Psi\left(x^{k+1}, t^{k+1}\right) &= \left\|\Phi\left(x^{k+1}, t^{k+1}\right)\right\| + \left\|t^{k+1} - F\left(x^{k+1}\right)\right\| \\ &\le \xi \max_{0\le r\le m(k)-1}\left(\left\|\Phi^{k-r}\right\| + \left\|t^{k-r} - F\left(x^{k-r}\right)\right\|\right) \\ &= \xi \max_{0\le r\le m(k)-1}\Psi\left(x^{k-r}, t^{k-r}\right) \\ &\le \Psi\left(x^0, t^0\right).\end{aligned} \tag{19}$$

So, $(x^{k+1}, t^{k+1}) \in L(x^0, t^0)$. Based on the similar analysis, it is easy to see $\{(x^k, t^k)\} \subset L(x^0, t^0)$ for all k. □

Theorem 2. *Assume Assumption 1 holds, and $\{(x^k, t^k)\}$ is generated by Algorithm1; then there exists an accumulation point (x^*, t^*) of the sequence $\{(x^k, t^k)\}$ which is solution of NCP(1).*

Proof. From Lemma 3 and Lemma 4, we know $\{(x^k, t^k)\} \subset L(x^0, t^0)$. By Assumption 1(b), we see that $L(x^0, t^0)$ is bounded. So, $\{(x^k, t^k)\}$ has an accumulation point. Suppose there exists a subsequence $\{(x^k, t^k)\}_{k\in K}$ which has an accumulation point (x^*, t^*). We should prove $H(x^*, t^*) = 0$.

Suppose $\{(x^k, t^k)\}_{k\in K}$ be an infinite sequence generated by Algorithm1. By construction of the algorithm, we know there are two types of successive iteration. Let $K_1 = \left\{k | x^{k+1} = x^k + d^k, t^{k+1} = t^k + \lambda^k\right\}$ and $K_2 = \{k | x^{k+1} = x^k + \rho_k d^k, t^{k+1} = t^k + \rho_k\lambda^k\}$. We need to prove the conclusion by the following two cases:

Case I: K_1 is an infinite index set. Let the sequence be $\{(x^k, t^k)\}_{k\in K_1}$, which satisfy (b) of Algorithm1. Therefore,

$$\Psi^{k_1} \le \xi\Psi^{k_2} \le \xi^2\Psi^{k_3} \le \cdots \le \xi^{m-1}\Psi^{k_m}. \tag{20}$$

This suggests that $\liminf_{k\longrightarrow\infty} H(x^k, t^k) = 0$.

Case II: K_2 is an infinite index set. Let the sequence be $\{(x^k, t^k)\}_{k\in K_2}$, which satisfy (f) and (g) of Algorithm 1.

It is known that $\|\Phi^{l(k)}\|$ is monotone decreasing and $\lim_{k\longrightarrow\infty}\|\Phi^k\| = 0$ by Lemma 2 and $\|t^{l(k)} - F(x^{l(k)})\|$ is

Step 0: initialization.
Given initial point $(x^0, t^0) \in R^{2n}$, $\mu \in (0,1)$, $\xi > 0$, $\bar{\xi} < 1$, $B_0 = V(x^0, t^0)$, $k = 0$.
Step 1: if $\Psi(x^k, t^k) = 0$, then stop. Otherwise, calculate the search direction

$$B_k \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} F(x^k) - t^k \\ -\Phi(x^k, t^k) \end{pmatrix}. \text{ (a)}$$

By (a), we can obtain d^k and λ^k.
Step 2: modified linear search technique.
Step 2.1 If

$$\Psi(x^k + d^k, t^k + \lambda^k) \le \xi \Psi(x^k, t^k), \text{ (b)}$$
$$\|\Psi(x^k + d^k, t^k + \lambda^k)\| \le \xi \max_{0 \le r \le m(k)-1} \|\Psi(x^{k-r}, t^{k-r})\|, \text{ (c)}$$
$$\|t^k + \lambda^k - F(x^k + d^k)\| \le \xi \max_{0 \le r \le m(k)-1} \|t^{k-r} - F(x^{k-r})\|, \text{ (d)}$$

where $m(0) = 0$, $0 \le m(k) \le \min\{m(k-1) + 1, M\}$ is a positive constant. Then, let

$$x^{k+1} = x^k + d^k, t^{k+1} = t^k + \lambda^k, \text{ (e)}$$

and go to Step 3; otherwise, go to Step 2.2.
Step 2.2: for $j = 0, 1, \ldots,$, check the following inequality with μ^j successively

$$\|\Phi(x^k + \mu^j d^k, t^k + \mu^j \lambda^k)\| \le \xi \max_{0 \le r \le m(k)-1} \|\Phi(x^{k-r}, x^{k-r})\|, \text{ (f)}$$
$$\|s^k + \mu^j \lambda^k - F(x^k + \mu^j d^k)\| \le \xi \max_{0 \le r \le m(k)-1} \|t^{k-r} - F(x^{k-r})\|, \text{ (g)}$$

Let j_k be the smallest nonnegative integer j such that (f) and (g) hold for μ^j. Set $\rho_k := \mu^{j_k}$, and

$$x^{k+1} = x^k + \rho_k d^k, t^{k+1} = t^k + \rho_k \lambda^k, \text{ (h)}$$

and go to Step 3.
Step 3: Update B_k to get B_{k+1},

$$B_{k+1} = B_k + \xi_k ((z^k)^T (y^k - B_k z^k) / \|z^k\|^2), \text{ (i)}$$

where

$$z^k = \begin{pmatrix} x^{k+1} \\ t^{k+1} \end{pmatrix} - \begin{pmatrix} x^k \\ t^k \end{pmatrix}, y^k = H(x^{k+1}, t^{k+1}) - H(x^k, t^k). \text{ (j)}$$

Select ξ_k to satisfy $|\xi_k - 1| \le \bar{\xi}$ and matrix B_{k+1} is nonsingular.
Step 4: Let $k = k + 1$, go to Step 1.

Algorithm 1

monotone decreasing and $\lim_{k \longrightarrow \infty} \|t^k - F(x^k)\| = 0$ by Lemma 3. Therefore, $\lim_{k \longrightarrow \infty} \|H(x^k, t^k)\| = \lim_{k \longrightarrow \infty} \|\Phi^k\| + \lim_{k \longrightarrow \infty} \|t^k - F(x^k)\| = 0$ as $k \in K_2$.

Therefore, the conclusion is followed. □

4. Numerical Results

In this section, some numerical results are given. We used a personal computer with 4.0 GB memory and Intel(R) Core(TM)i5-5200U CPU @2.20 GHz to perform all experiments. We used Windows 10 as the operating system and Matlab R2018b to write the computer codes. In the whole experiment, the parameters used in Algorithm1 were $\xi = 0.9$, $\mu = 0.8$, $\xi_k \equiv 1$, M is an integer which is randomly selected from 2 to 5. $\|H(x,t)\| < 10^{-6}$ was the stop criterion.

The number of iterations, the CPU time in seconds, and the value of $x^{(T)} F(x)$ at the final iteration are are listed in Table 1. x_0 in Table 1 means the Initial point where ones $(i, 1)$ means the i dimension of this problem.

4.1. Some Test Problems. Examples 1–8 (NCP) are considered.

Example 1. Consider (1), where $x \in R^n$and $F(x) = Mx + q$ with

$$M = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 4 \end{pmatrix},$$
$$q = \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ -1 \end{pmatrix}. \tag{21}$$

We use $x^0 = (1, 1, \ldots, 1)^T$and $t^0 = (10^{-3}, 10^{-3}, \ldots, 10^{-3})^T$as the starting points to text this problem.

Example 2. Consider (1), where $x \in R^3$, and $F(x)$: $R^3 \longrightarrow R^3$ given by

$$F(x) = \begin{pmatrix} x_2 \\ x_3 \\ -x_2 + x_3 + 1 \end{pmatrix}. \tag{22}$$

Figure 1 is the 3D diagram of Example 2. $(0, \lambda, 0)$ is an infinite solution of this problem, where $\lambda \in [0, 1]$. The initial points x^0, t^0 are randomly generated, and these elements are in the interval $(0, 10)$.

Example 3. Consider (1), where $x \in R^7$, and $F(x)\colon R^7 \longrightarrow R^7$ given by

$$F(x) = \begin{pmatrix} 2x_1 - x_3 + x_5 + 3x_6 - 1 \\ x_2 + 2x_5 + x_6 - x_7 - 3 \\ -x_1 + 2x_3 + x_4 + x_5 + 2x_6 - 4x_7 + 1 \\ x_3 + x_4 + x_5 - x_6 - 1 \\ -x_1 - 2x_2 - x_3 - x_4 + 5 \\ -3x_1 - x_2 - 2x_3 + x_4 + 4 \\ x_2 + 4x_3 - 1.5 \end{pmatrix}. \quad (23)$$

Example 4. Consider (1), where $x \in R^4$ and $F(x)\colon R^4 \longrightarrow R^4$ given by

$$F(x) = \begin{pmatrix} x_1^3 - 8 \\ x_2 + x_2^3 - x_3 + 3 \\ x_2 + x_3 + 2x_3^3 - 3 \\ x_4 + 2x_4^3 \end{pmatrix}. \quad (24)$$

Example 5 (Kojima–Shindo Problem). Consider (1), where $x \in R^4$and $F(x)\colon R^4 \longrightarrow R^4$ given by

$$F(x) = \begin{pmatrix} 3x_1^2 + 2x_1x_2 + 2x_2^2 + x_3 + 3x_4 - 6 \\ 2x_1^2 + x_1 + x_2^2 + 10x_3 + 2x_4 - 2 \\ 3x_1^2 + x_1x_2 + 2x_2^2 + 2x_3 + 9x_4 - 9 \\ x_1^2 + 3x_2^2 + 2x_3 + 3x_4 - 3 \end{pmatrix}. \quad (25)$$

$(\sqrt{6}/2, t\,n0q, h_0x, 7\,C0.5)$ is a degenerate solution, and $(1, 0, 3, 0)$ is a nondegenerate solution.

Example 6 (Modified Mathiesen Problem). Consider (1), where $x \in R^4$ and $F(x)\colon R^4 \longrightarrow R^4$ given by

$$F(x) = \begin{pmatrix} -x_2 + x_3 + x_4 \\ x_1 - \dfrac{4.5x_3 + 2.7x_4}{x_2 + 1} \\ 5 - x_1 - \dfrac{0.5x_3 + 0.3x_4}{x_3 + 1} \\ 3 - x_1 \end{pmatrix}. \quad (26)$$

Example 7. The function $f(x)$ is endowed with the component as follows:

$$\begin{aligned} F(x) &= \left(f_1(x), f_2(x), \ldots, f_n(x)\right)^T, \\ f_i(x) &= e^{x_i} - 1, \quad i = 1, 2, \ldots, n-1, \\ f_n(x) &= e^{x_n} + x_n - 1. \end{aligned} \quad (27)$$

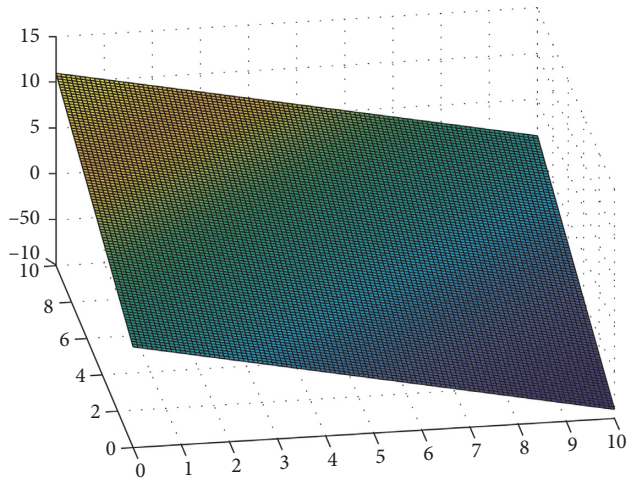


Figure 1: Diagram of Example 2.

Example 8. Consider (1), where $x \in R^n$ and $F(x) = Mx + q$ with

$$M = \operatorname{diag}\left(\frac{1}{n}, \frac{2}{n}, \ldots, 1\right), q = (-1, -1, \ldots, -1)^T. \quad (28)$$

Table 1 shows the results of Examples 1–8 using 3-1 piecewise, 4-1 piecewise Algorithm1 and feasible direction method, respectively. It can be seen from the table that Algorithm1 applying 3-1 piecewise has a good solution to all the above problems. Algorithm1 applying 4-1 piecewise is slightly insufficient, and the feasible direction method has some difficulties in solving examples above, and some of the examples cannot be solved. Figure 2 shows how the $x^T f(x)$ value of the three algorithms decreases as the number of iterations increases in each specific example. We use performance profiles [28]—distribution functions for a performance metric—as a tool for comparing different algorithms. We consider the comprehensive performance of the above three algorithms in terms of CPU time, number of iterations, and $x^T f(x)$ value. If the curve is closer to 1, the better the ability to solve the problem (Figure 3).

4.2. Nash Equilibrium Problem. General economic equilibrium [29] means that total supply and total demand are exactly equal in a price system. With the existing productivity and technical conditions, producers get the most profit, while consumers get the most utility when they meet the budget constraints. The theory of general economic equilibrium was first put forward by the French economist Walras. Walras believes that when the whole economy is in equilibrium, the prices of all consumer goods and factors of production will have a certain equilibrium value, and their output and supply will have a certain equilibrium quantity. It is assumed that the whole economic system is a large and complete trading market, and the equilibrium price system means that all commodities are traded in this market, and finally all commodities can be traded.

Table 1: Iterations, CPU time, and $x^T f(x)$ for NCP Examples 2–6 between Algorithm1 and FDA.

Problem	x^0	Algorithm 1 with 3–1 piecewise			Algorithm 1 with 4–1 piecewise			Feasible directions algorithm		
		Iter	CPU time	$x^T f(x)$	Iter	CPU time	$x^T f(x)$	Iter	CPU time	$x^T f(x)$
4.1	*ones*(100, 1)	2	0.000814	$5.49E-09$	13	0.001768	$1.12E-08$	5	0.004301	$6.58E-07$
		2	0.473808	$5.69E-08$	13	1.520458	$1.21E-08$	5	0.257754	$-8.81E-14$
	ones(4069, 1)	3	15.640096	$-1.58E-11$	13	18.551757	$1.16E-08$	5	8.556959	$7.20E-14$
	ones(8138, 1)	2	51.204458	$4.51E-10$	13	70.145272	$1.44E-08$	5	53.324527	$-1.27E-12$
4.2	$(1,1,1)^T$	3	0.000235	$-1.02E-07$	8	0.000617	$-5.41E-08$	10	0.000354	$4.66E-07$
	$(10,10,10)^T$	4	0.000288	$5.25E-08$	10	0.000604	$3.26E-07$	13	0.000529	$3.76E-07$
	$(1,5,9)^T$	7	0.000462	$-2.23E-08$	5	0.000742	$4.64E-07$	19	0.001465	$3.50E-07$
	ones(7, 1)	15	0.001014	$-2.23E-08$	22	0.001768	$5.53E-07$	26	0.004301	$6.58E-07$
4.3	10 ∗ ones(7, 1)	15	0.001661	$-4.62E-09$	24	0.002366	$1.05E-08$	NaN	NaN	NaN
	10 ∗ rand(7, 1)	16	0.001608	$-2.14E-09$	20	0.001204	$1.20E-08$	NaN	NaN	NaN
4.4	$(1,1,1,1)^T$	14	0.007181	$-3.11E-08$	23	0.001696	$-6.07E-07$	>500∗	Inf	NaN
	$(10,10,10,10)^T$	145	0.019268	$1.80E-08$	86	0.004996	$-2.26E-07$	21	0.007137	$6.12E-07$
4.5	$(1,1,1,1)^T$	13	0.00103	$-5.62E-09$	26	0.001658	$4.93E-09$	36	0.001121	$6.43E-07$
	$(1,2,3,4)^T$	23	0.002497	$-5.65E-07$	30	0.003223	$-2.75E-08$	65	0.009764	$9.61E-07$
4.6	$(1,1,1,1)^T$	10	0.001351	$1.52E-08$	9	0.001473	$6.19E-09$	16	0.00746	$3.67E-07$
4.7	ones(100, 1)	11	0.039089	$-9.29E-11$	17	0.001768	$8.98E-08$	46	0.024301	$8.81E-07$
	ones(1024, 1)	20	1.230144	$2.26E-10$	20	1.745354	$-2.68E-11$	52	1.157443	$7.57E-07$
	ones(4069, 1)	15	25.532848	$-2.61E-12$	20	35.615797	$3.61E-10$	55	50.806954	$8.14E-07$
	ones(8138, 1)	15	305.640096	$-4.09E-14$	20	370.126146	$4.62E-09$	57	415.720027	$7.56E-07$
4.8	ones(100, 1)	4	0.002064	$7.41E-07$	57	0.107685	$1.32E-07$	NaN	NaN	NaN
	ones(1024, 1)	6	0.641349	$5.40E-07$	407	22.002366	$3.10E-06$	NaN	NaN	NaN
	ones(4069, 1)	8	26.727794	$1.87E-06$	>500*	Inf	NaN	NaN	NaN	NaN
	ones(8138, 1)	10	215.640096	$5.19E-06$	>500*	Inf	NaN	NaN	NaN	NaN

Considering the competitive economic model of production and investment, suppose H is a price system, in which there are N kinds of commodities, we use R^N to express commodity space. For producer i, the set of production is $Y_i \subseteq R^N$. For consumer j, the set of consumption is $Z_j \subseteq R^N$. The number of producers and consumers in the system are l and k, respectively. The total production, total consumption, and initial commodity reserve are represented by Y_i, Z_j, and λ_j, respectively, and the proportion of consumer j in the profit of producer i is represented by ϕ_{ji}. Specially, $i = 1, \ldots, l$; $j = 1, \ldots, k$; and Z_j, Y_i, $\lambda_j \in R^N$.

To describe the model better, we assume the following definitions. In particular, Z_j, Y_i, and λ_j are independent of x.

Definition 1. Let $z_j \in Z_j$, $y_i \in Y_i$, x is the equilibrium price:

(1) For every i, the maximum profit function is $x \cdot y_i$.

(2) For every j, preference maximum element is $z_j = \{z_j \in Z_j | xt \cdot nz_j q \le hx.x\lambda_j 7 + C \sum_{j=1}^{l} \phi_{ji} \cdot x \cdot y_i$.

(3) Economic equilibrium is defined as $\sum_{i=1}^{l} \lambda_i + \sum_{i=1}^{l} x \cdot y_i - \sum_{j=1}^{k} z_j = 0$.

It can be seen from Definition 1 that when price system H reaches economic equilibrium, the demands of both producers and consumers are satisfied and then all the commodities of price system H are sold, that is, the commodities are cleared. We define the conditions for clearing the goods as

$$F = \sum_{i=1}^{l} \lambda_i + \sum_{i=1}^{l} x \cdot y_i - \sum_{j=1}^{k} z_j, \qquad x \ge 0, \ x \cdot F = 0. \tag{29}$$

Equation (29) is not only the equilibrium state of free allocation, but also the model of linear complementarity problem. If Z_j, Y_i, and λ_j are related to x, (29) will become a nonlinear complementarity problem (NCP).

Let the inverse demand function for the market be defined by

$$P(Q) = 5000^{(1/\gamma)} Q^{-(1/\gamma)}, \tag{30}$$

where Q is the total quantity produced, P is the market price, and γ is the elasticity of demand with respect to price. Let q_i denote the output of firm i and let the total cost function for firm i be given by

$$\begin{aligned} f_i(q_i) &= c_i q_i + \left(\frac{\beta_i}{1+\beta_i}\right) L^{(1/\beta_i)} q_i^{((\beta_i+1)/\beta_i)}, \\ F_i(q) &= f_i'(q_i) - p \sum_{j=1}^{n} q_j - q_i p\prime \sum_{j=1}^{n} q_j, \quad i = 1, 2, \ldots, n, \\ F &= [F_1(q), F_2(q), \ldots, F_i(q)]. \end{aligned} \tag{31}$$

Example 9. Data is given in Table 2.

Example 10. Data is given in Table 3.

4.3. Two-Dimensional Contact Problem. Under the conditions of nonpenetration and negligible attraction between objects, the elastic contact problem mainly requires the

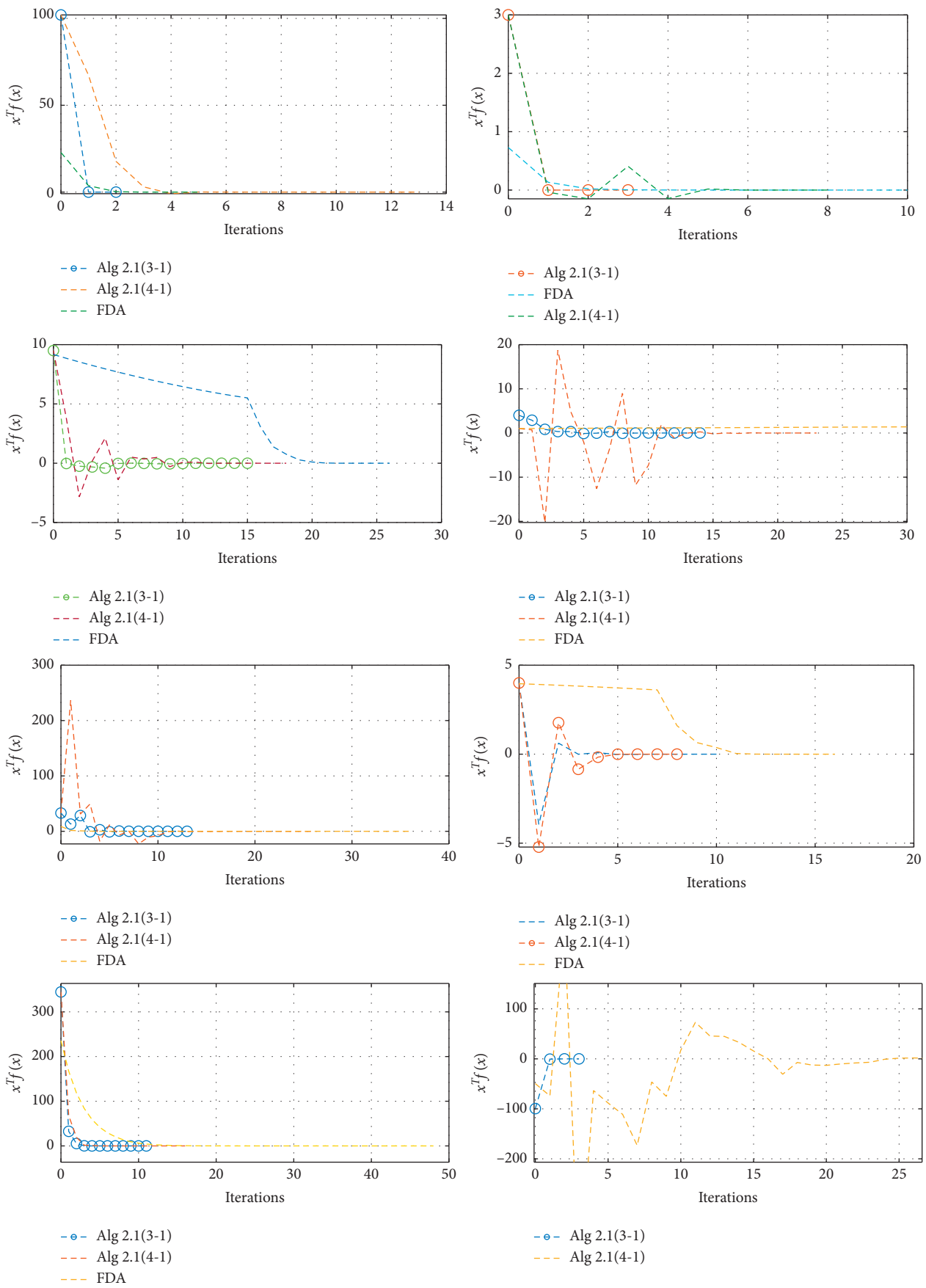


Figure 2: Schematic diagram of the changes of $x^T f(x)$ with iteration of the three algorithms (the same initial point of ones $(n, 1)$).

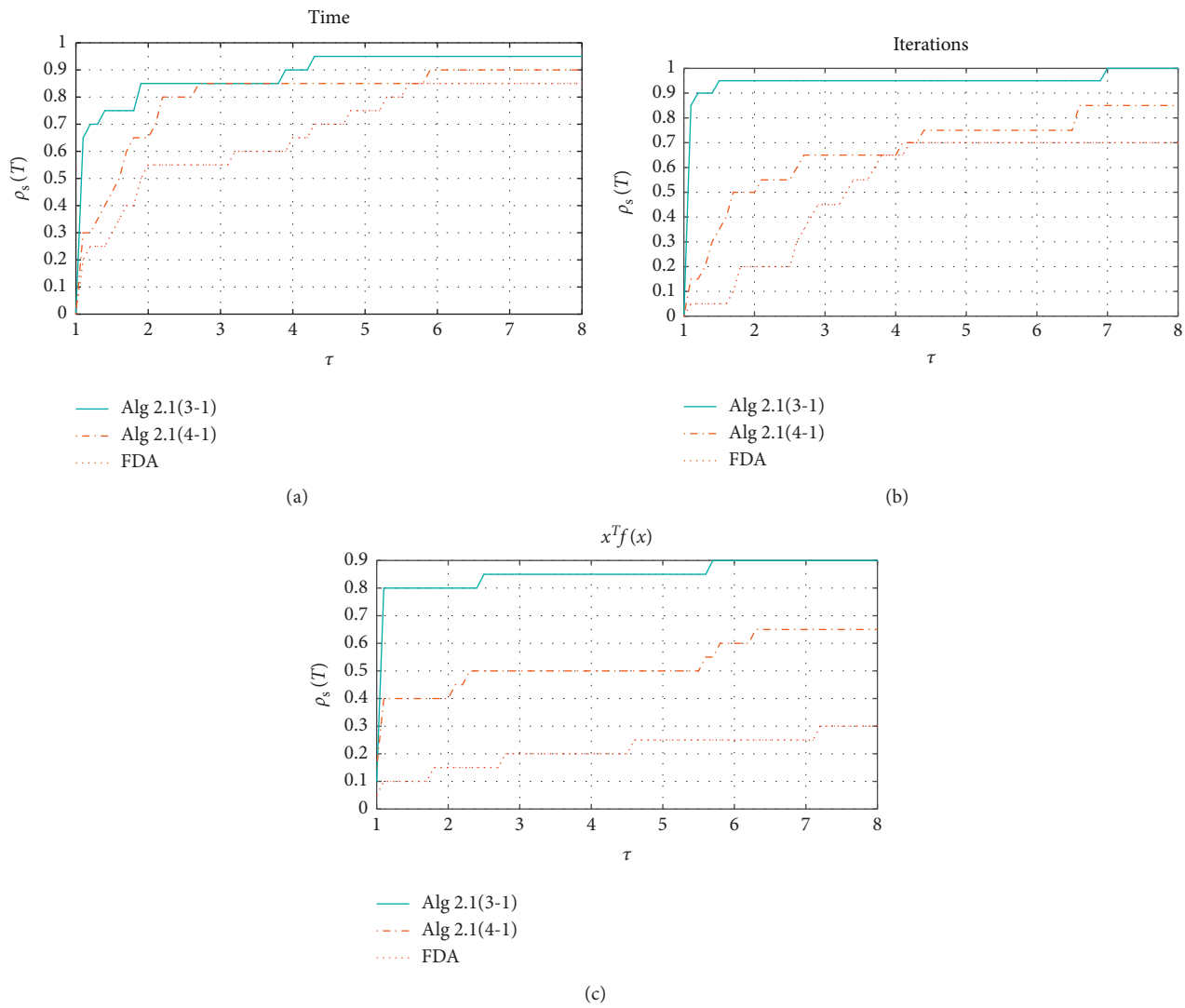


Figure 3: Performance profile for Algorithm1 and feasible directions algorithm through Examples 1–8.

contact surface and the pressure of the contact surface when two objects are pressed together. The wheel-rail problem is a typical elastic contact problem.

Figure 4 [31] shows the geometric structure application of the wheel-rail contact phenomenon, where Figure 4(a) represents the overall geometric structure showing the forward speed V and angular velocity ω of the track when the wheel is rolling. The track is deformed by the wheel pressure Fw and the sleeper pressure Fs1 and Fs2. At the same time, the wheel deforms due to the wheel-rail pressure Fr, and Figures 4(b) and 4(c) represent the undeformed and deformed states, respectively.

Regarding a point (x, y) on the contact surface, if z represents the pressure on the point, u represents the displacement from the dashed line to the solid line along the normal direction, q represents the distance of the dashed line when the point is not deformed, and w represents its shape, the gap between the rear wheel and the track is $w = u + q$. Assume that C is the contact surface and E is the other external area, the geometric relationship shown in Figure 4 can be abbreviated as

$$\begin{aligned} &\forall (x, y) \in C, \quad w = 0, \ z \geq 0, \\ &\forall (x, y) \in E, \quad w > 0, \ z = 0. \end{aligned} \tag{32}$$

If the two-dimensional potential contact area with contact surface is discretized, users $mx \times my$ grid is divided, and let n represent the total number of grids; then

$$u = Tz, z, u \in R^n, T \in R^{n \times n}, \tag{33}$$

and the problem can be changed into a linear complementarity problem LCP(q, T); to find a pair $w, z \in R^n$, the following is satisfied

$$w = Tz + q \geq 0, z \geq 0, z^T w = 0, \tag{34}$$

where the coefficient matrix [32] T is a Toeplitz matrix, satisfying

Table 2: Data of Example 9.

Firm i	c_i	L_i	β_i
1	10	5	1.2
2	8	5	1.1
3	6	5	1
4	4	5	0.9
5	2	5	0.8

Table 3: Data of Example 10.

Firm i	c_i	L_i	β_i
1	5	10	1.20
2	3	10	1.00
3	8	10	0.90
4	5	10	0.60
5	1	10	1.50
6	3	10	1.00
7	7	10	0.70
8	4	10	1.10
9	6	10	0.95
10	3	10	0.75

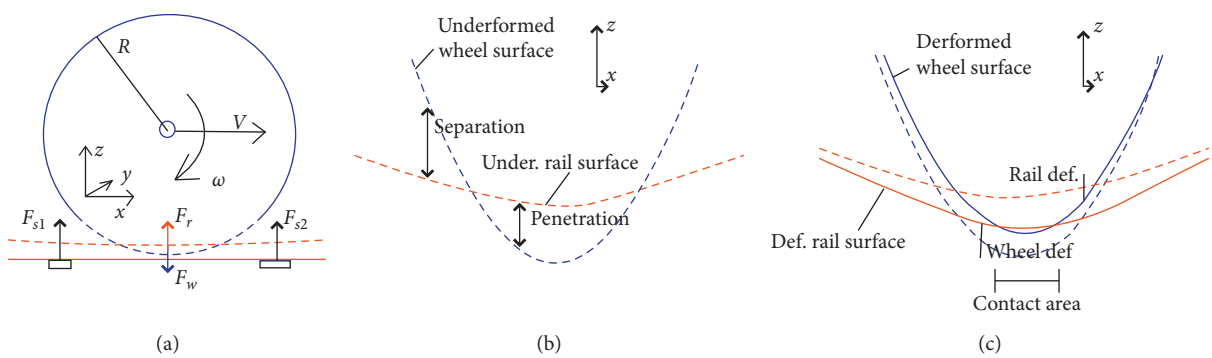


Figure 4: Schematic diagram of the two-dimensional contact problem [30].

$$T = \begin{pmatrix} T_0 & T_{-1} & \cdots & T_{2-n} & T_{1-n} \\ T_1 & T_0 & T_{-1} & \cdots & T_{2-n} \\ \vdots & T_1 & T_0 & \ddots & \vdots \\ T_{n-2} & \vdots & \ddots & \ddots & T_{-1} \\ T_{n-1} & T_{n-2} & \cdots & T_1 & T_0 \end{pmatrix}. \tag{35}$$

Example 11. The diagonal element T_k of the coefficient matrix T is

$$T_k = \begin{cases} 2(1+k)^{-1.2}, & k \neq 0; \\ 2, & k = 0. \end{cases} \tag{36}$$

Example 12. The diagonal element T_k of the coefficient matrix T is

$$T_k = 2^{-k}, \quad k = 0, 1, \ldots, n-1. \tag{37}$$

Example 13. The diagonal element T_k of the coefficient matrix T is

$$T_k = \begin{cases} \left(\frac{19}{8}\right) + \left(\frac{1}{n}\right), & k = 0; \\ -0.5, & k = 1; \\ 0.25, & k = 2; \\ \left(\frac{1}{16}\right), & k = 3; \\ 0, & \text{else}. \end{cases} \tag{38}$$

Table 4 shows the performance of Algorithm1 using different piecewise methods for practical application problems. From Figures 5 to 7, it can be seen that Algorithm1 using 3-1 piecewise has a stronger ability to solve all the above problems than Algorithm1 applying 4-1 piecewise.

Table 4: Iterations, CPU time, and $x^T f(x)$ for problem 4.9–4.13.

Problem	x0	Algorithm 1 with 3–1 piecewise			Algorithm 1 with 4–1 piecewise		
		Iter	CPU time	XTF(X)	Iter	CPU time	XTF(X)
4.9	$20 * \text{ones}(5,1)$	20	0.005197	$-1.87E-06$	23	0.00523	$-1.49E-06$
	$30 * \text{ones}(5,1)$	18	0.010009	$-2.28E-06$	21	0.019381	$-3.98E-06$
4.10	$20 * \text{ones}(5,1)$	45	0.033759	$3.61E-06$	43	0.050365	$1.03E-06$
	$30 * \text{ones}(5,1)$	55	0.05047	$7.62E-07$	49	0.055852	$-1.80E-06$
4.11	$\text{ones}(100,1)$	13	0.052309	$1.87E-08$	15	0.089898	$1.65E-08$
	$\text{ones}(1024,1)$	14	2.080973	$-3.66E-09$	16	2.616348	$1.94E-09$
	$\text{ones}(4069,1)$	14	60.367315	$1.05E-09$	16	68.659795	$8.54E-09$
	$\text{ones}(8138,1)$	14	574.245627	$3.05E-09$	20	682.960276	$-5.23E-09$
4.12	$\text{ones}(100,1)$	14	0.058579	$1.96E-09$	14	0.072436	$-5.59E-09$
	$\text{ones}(1024,1)$	14	2.376266	$-1.89E-08$	14	2.348159	$-4.86E-09$
	$\text{ones}(4069,1)$	12	53.195688	$-7.32E-09$	14	61.321038	$2.96E-12$
	$\text{ones}(8138,1)$	17	519.846528	$-2.89E-08$	14	582.933657	$3.25E-11$
4.13	$\text{ones}(100,1)$	17	0.060449	$1.99E-08$	19	0.0813907	$4.00E-09$
	$\text{ones}(1024,1)$	17	2.860266	$5.62E-08$	19	3.145843	$8.61E-09$
	$\text{ones}(4069,1)$	17	74.532155	$7.27E-08$	19	82.960276	$-2.45E-09$
	$\text{ones}(8138,1)$	21	674.226819	$2.02E-08$	23	782.155632	$5.12E-09$

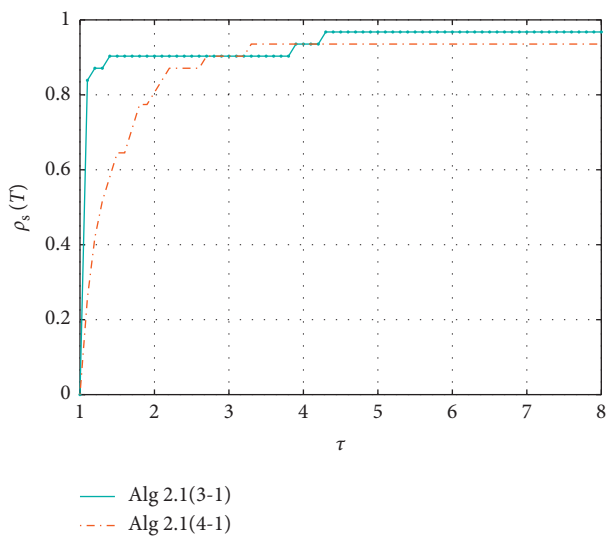


Figure 5: Performance profile on CPU time for Algorithm1 with different piecewise functions.

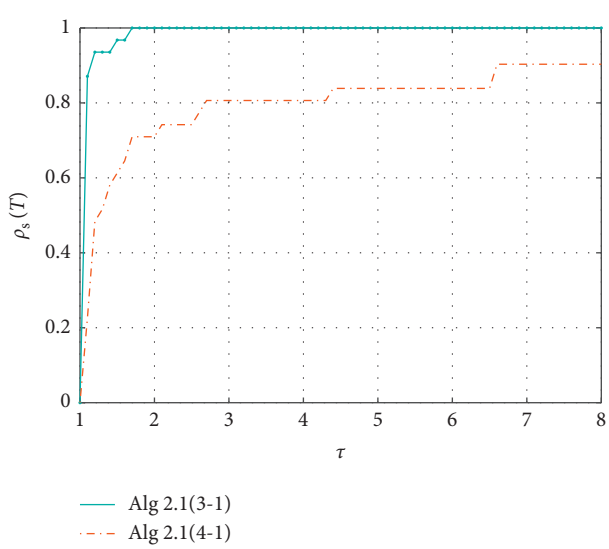


Figure 6: Performance profile on iterations for Algorithm1 with different piecewise functions.

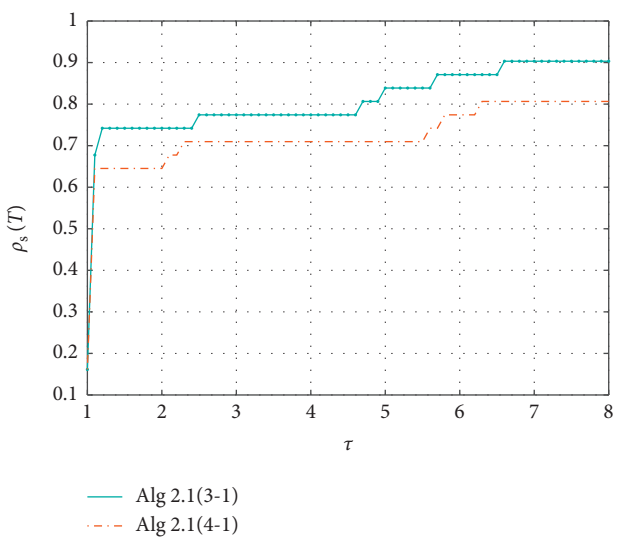


Figure 7: Performance profile on $x^T f(x)$ for Algorithm1 with different piecewise functions.

5. Conclusion

In this paper, by using 3-1 and 4-1 piecewise nonlinear complementarity problem functions, we reformulate the nonlinear complementarity problem into smooth equations. By using a new nonmonotone line search, a modified smooth Broyden-like algorithm is proposed and the global convergence of the proposed algorithm is obtained, and the numerical tests for some practical problems show the efficiency of the algorithm. How to get the local convergence under certain conditions is worth studying in the future.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interests regarding the publication of this paper.

References

[1] P. T. Harker and J.-S. Pang, "Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications," *Mathematical Programming*, vol. 48, no. 1–3, pp. 161–220, 1990.

[2] J. J. Moré, "Global methods for nonlinear complementarity problems," *Mathematics of Operations Research*, vol. 21, no. 3, pp. 589–614, 1996.

[3] M. C. Ferris and J. S. Pang, "Engineering and economic applications of complementarity problems," *Society of Indian Automobile Manufacturers Review*, vol. 39, no. 4, pp. 669–713, 1997.

[4] A. Fischer, "A special Newton-type optimization method," *Optimization*, vol. 24, no. 3-4, pp. 269–284, 1992.

[5] B. Chen, X. Chen, and C. Kanzow, "A penalized Fischer-Burmeister NCP-function," *Mathematical Programming*, vol. 88, no. 1, pp. 211–216, 2000.

[6] D. Sun and L. Qi, "On NCP-functions," *Computational Optimization and Applications*, vol. 13, no. 1–3, pp. 201–220, 1999.

[7] T. Luca, F. Facchinei, and C. Kanzow, "A semismooth equation approach to the solution of nonlinear complementarity problems," *Mathematical Programming*, vol. 75, no. 3, pp. 407–439, 1996.

[8] H. Jiang and L. Qi, "A new nonsmooth equations approach to nonlinear complementarity problems," *SIAM Journal on Control and Optimization*, vol. 35, no. 1, Article ID 178193, 1997.

[9] T. Luca, F. Facchinei, and C. Kanzow, "A theoretical and numerical comparison of some semismooth algorithms for complementarity problems," *Computational Optimization and Applications*, vol. 16, pp. 173–205, 2000.

[10] L. Qi and Y. Yang, "NCP functions applied to Lagrangian globalization for the nonlinear complementarity problem," *Journal of Global Optimization*, vol. 24, Article ID 261283, 2002.

[11] J.-S. Pang and S. A. Gabriel, "NE/SQP: a robust algorithm for the nonlinear complementarity problem," *Mathematical Programming*, vol. 60, no. 1–3, pp. 295–337, 1993.

[12] C. Chen and O. L. Mangasarian, "A class of smoothing functions for nonlinear complementarity problems," *Journal of Computational and Applied Mathematics*, vol. 80, pp. 105–126, 1997.

[13] C. Kanzow and H. Pieper, "Jacobian smoothing methods for nonlinear complementarity problems," *SIAM Journal on Optimization*, vol. 9, no. 2, pp. 342–373, 1999.

[14] N. Krejic and S. Rapaji´ c, "Globally convergent Jacobian smoothing inexact´ Newton methods for NCP," *Computational Optimization and Applications*, vol. 41, pp. 243–261, 2008.

[15] L. Qi and D. H. Li, "A smoothing Newton method for nonlinear complementarity problems," *Advanced Modeling and Optimization*, vol. 13, no. 2, pp. 141–152, 2011.

[16] A. Liu, D. G. Pu, and D. Pu, "3-1 piecewise NCP function for new nonmonotone QP-free infeasible method," *Journal of Robotics and Mechatronics*, vol. 26, no. 5, pp. 566–572, 2014.

[17] S. Ke and D. Yang, "A smooth Newton method with 3-1 piecewise NCP function for generalized nonlinear complementarity problem," *International Journal of Computer Mathematics*, vol. 95, pp. 1703–1713, 2018.

[18] S. Ke and D. Yang, "A modified non-monotone method with 3-1 piecewise NCP function for nonlinear complementary problem," *Computer Model. New Tech.*vol. 21, no. 1, pp. 47–51, 2017.

[19] L. Qi and D. Sun, "Smoothing functions and smoothing Newton method for complementarity and variational inequality problems," *Journal of Optimization Theory and Applications*, vol. 113, no. 1, pp. 121–147, 2002.

[20] L. Qi, D. Sun, and G. Zhou, "A new look at Smoothing Newton method for complementarity problems and box constrained variational inequality problems," *Antimicrobial resistance*, vol. 97, p. 13, 1997.

[21] X. Chen, L. Qi, and D. Sun, "Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities," *Mathematics of Computation*, vol. 67, no. 222, pp. 519–541, 1998.

[22] J. Tang, L. Dong, J. Zhou, and L. Fang, "A smoothing Newton method for nonlinear complementarity problems," *Computational and Applied Mathematics*, vol. 32, no. 1, pp. 107–118, 2013.

[23] C. Ma, X. Chen, and J. Tang, "On convergence of a smoothing Broyden-like method for -NCP," *Nonlinear Analysis: Real World Applications*, vol. 9, no. 3, pp. 899–911, 2008.

[24] B. Chen and C. Ma, "A new smoothing Broyden-like method for solving nonlinear complementarity problem with a P 0-function," *Journal of Global Optimization*, vol. 51, no. 3, pp. 473–495, 2011.

[25] B. Fan, "A smoothing Broyden-like method with a nonmonotone derivative-free line search for nonlinear complementarity problems," *Journal of Computational and Applied Mathematics*, vol. 290, pp. 641–655, 2015.

[26] X. Y. Zheng, J. R. Shi, W. Yang, and Q. Y. Yin, "Nonmonotone smoothing Broyden-like method for generalized nonlinear complementarity problems," *Journal of Applied Mathematics and Computing*, vol. 26, pp. 566–572, 2014.

[27] J. S. Chen and S. H. Pan, "A regularization semismooth Newton method based on the generalized fischer-burmeister function for P0 -NCPs," *Journal of Computational and Applied Mathematics*, vol. 1, pp. 464–476, 2007.

[28] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical Programming*, vol. 91, no. 2, pp. 201–213, 2002.

[29] L.-P. Zhang and Y. Zhou, "A note on economic equilibrium and financial networks," *Acta Mathematicae Applicatae Sinica, English Series*, vol. 30, no. 1, pp. 89–98, 2014.

[30] J. Zhao, E. A. H. Vollebregt, and C. W. Oosterlee, "A full multigrid method for linear complementarity problems arising from elastic normal contact problems," *Mathematical Modelling and Analysis*, vol. 19, no. 2, pp. 216–240, 2014.

[31] J. Zhao, E. A. H. Vollebregt, and C. W. Oosterlee, "A full multigrid method for linear complementarity problems arising from elastic normal contact problems," *Mathematical Modelling and Analysis*, vol. 19, no. 2, pp. 216–240, 2014.

[32] M. H. Wu and C. L. Li, "A preconditioned modulus-based matrix multisplitting block iteration method for the linear complementarity problems with Toeplitz matrix," *Calcolo*, vol. 56, p. 2, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9507983, 11 pages
https://doi.org/10.1155/2020/9507983

Research Article

Low-Speed Stability Optimization of Full-Order Observer for Induction Motor

Xiangsheng Liu,[1] **Lin Ren**,[2] **Yuanyuan Yang**,[1] **Jun He**,[1] **and Zhengxin Zhou**[1]

[1]*College of Mechanical and Electrical Engineering, Sanda University, Shanghai 201209, China*
[2]*PingDingshan Vocational and Technical College, Henan 467000, China*

Correspondence should be addressed to Lin Ren; renlin_chn@163.com

Received 16 August 2020; Revised 8 October 2020; Accepted 6 November 2020; Published 11 December 2020

Academic Editor: Mohamed El Ghami

In terms of the instability of the full-order observer for the induction motor in the low-speed regenerative mode, the low-speed unstable region which leads to the extension of the commissioning cycle cannot be eliminated by the traditional adaptive law which aims at good system performance. It is proposed that the feedback gain matrix can control both the unstable region and the system performance both. To make a trade-off between the stability and performance by designing the feedback gain matrix is still an open problem. To solve this problem, first we analyze the cause of instability and derive constraints to ensure system stability by establishing a transfer function of the adaptive observing system for the speed. Then, with the derived constraints as the design criteria for the feedback gain matrix, a control strategy combining the weighted adaptive law with the improved feedback gain matrix is proposed to improve the stability at low speed. Finally, by comparing the traditional control strategy with the proposed control strategy through simulations and experiments, we show that the proposed control strategy achieves better performance with higher stability.

1. Introduction

The speed-sensorless vector control system of the induction motor abandons the photoelectric encoder and other traditional motor speed measurement devices, which reduces the cost of the system and enhances the reliability of system operation. At present, among speed identification methods for the speed-sensorless induction motor, the direct calculation method [1] directly uses the mathematical model of the induction motor for speed open-loop estimation. Although the structure is simple, this method features poor anti-interference ability and low-speed identification accuracy. The model reference adaptive control method [2, 3] takes the voltage model as an adjustable one that has a simple principle. However, the pure integrator in the voltage model causes DC bias and error in integral initial value, which leads to poor performance at low speed. The high frequency signal injection method [4] eliminates the problem of poor low-speed performance of the model reference adaptive control method by taking advantage of the salient pole rotors. However, it depends heavily on the structural design of the motor and is not practical enough. In the adaptive full-order observer method [5], a state equation of the rotor-flux linkage and the stator current is established to predict the state of the motor in real time for the induction motor. The difference between the estimated value and the measured value of the stator current state is corrected and input by the gain matrix, and the estimated state is corrected in real time by feedback correction, thus forming a closed-loop state estimation to improve the performance of the speed identification system.

As a widely used speed identification tool, the adaptive full-order observer is unstable in the low-speed regenerative mode. To address this problem, there are many works on improving the speed identification system. In References [5, 6], the rotor-flux linkage error is ignored in the process of deriving the speed adaptive law using Popov's hyperstability theory. Although the immeasurability of the rotor-flux

linkage is considered, when the motor runs at low speed, the rotor-flux linkage error increases significantly, which results in inaccurate speed identification. In Reference [7], the rotor-flux linkage error is compensated in the adaptive law, which improves the accuracy and dynamic performance of the speed identification system. However, in the design of the weight coefficient of the rotor-flux linkage error in the scheme, filtering processing is required, which leads to an increase in system complexity. Since the poles of the motor model are in the left half plane of the s-plane, the model itself is stable [8]. In Reference [9], it is proposed that the poles of the full-order observer should be set on the left side of the motor pole. The scheme can improve the convergence speed of the full-order observer to a certain extent by setting a reasonable feedback gain matrix. However, the stability of the low-speed regenerative mode is still not effectively solved. In Reference [10], the transfer function of the open-loop full-order observer is analyzed, and the unstable region under the low-speed regenerative mode is given. In Reference [11], the regenerative instability problem is solved by improving the feedback gain matrix, but the pole position of the full-order observer is moved to the position close to the origin, which reduces the convergence speed of the system. References [12–17] provide a new idea for speed-sensorless performance optimization at low speed, but its algorithm is not practical due to its complexity.

In view of the shortcomings of the improved adaptive law [5–7] and the feedback gain matrix [8–11] of the full-order observer, an improved method combining the adaptive law with the feedback gain matrix is proposed to improve the dynamic performance and low-speed stability of the system, by introducing an adaptive law compensation method with adjustable weight coefficient and simplifying the feedback gain matrix with low-speed stability as the design criteria. The feasibility and effectiveness of this control strategy are supported by theoretical analyses and simulations.

2. Mathematical Model of Full-Order Observer for Induction Motor

With stator current and rotor-flux linkage of the induction motor as state variables, the state equation of the induction motor in the static coordinate system is given by

$$\begin{cases} \dfrac{d}{\mathrm{d}t}x = Ax + Bu_s, \\ y = Cx. \end{cases} \tag{1}$$

By formula (1), the state equation of the full-order observer is obtained as follows:

$$\begin{cases} \dfrac{d}{\mathrm{d}t}\hat{x} = \hat{A}\hat{x} + Bu_s + G(\hat{y} - y), \\ \hat{y} = C\hat{x}, \end{cases} \tag{2}$$

where $C = \begin{bmatrix} I & 0 \end{bmatrix}$ is the output matrix, and the feedback gain matrix is as follows:

$$\begin{aligned} G &= \begin{bmatrix} G_1 & G_2 \end{bmatrix}^T = \begin{bmatrix} g_1 I + g_2 J & g_3 I + g_4 J \end{bmatrix}^T, \\ A &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} a_{11} I & a_{12} I + a_{12}' J \\ a_{21} I & a_{22} I + a_{22}' J \end{bmatrix} \\ &= \begin{bmatrix} \left(\dfrac{\delta - 1}{\delta T_r} - \dfrac{R_s}{\delta L_s}\right) I & \dfrac{L_m}{\delta L_s L_r T_r} I - \dfrac{L_m \omega_r}{\delta L_s L_r} J \\ \dfrac{L_m}{T_r} I & -\dfrac{1}{T_r} I + \omega_r J \end{bmatrix}, \end{aligned} \tag{3}$$

in which $B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$, $B_1 = (1/\delta L_s) I$, $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, $\delta = 1 - (L_m^2/L_s\ L_r)$, $T_r = (L_r/R_r)$, and $x = \begin{bmatrix} i_s & \psi_r \end{bmatrix}^T$ are the state variables, $y = i_s$ is the output variable, $i_s = \begin{bmatrix} i_{s\alpha} & i_{s\beta} \end{bmatrix}^T$ is the stator current, $\psi_r = \begin{bmatrix} \psi_{r\alpha} & \psi_{r\beta} \end{bmatrix}^T$ is the rotor-flux linkage, $u_s = \begin{bmatrix} u_{s\alpha} & u_{s\beta} \end{bmatrix}^T$ is the stator voltage, ω_r is the rotor speed, R_r and R_s are the rotor resistance and stator resistance, L_r, L_s, and L_m are the rotor inductance, stator inductance, and mutual inductance. The superscript "^" indicates the observed value.

An error equation is obtained by subtracting the state equation (1) of the induction motor from the state equation (2) of the full-order observer as follows:

$$\frac{d}{\mathrm{d}t}\begin{bmatrix} e_i \\ e_\psi \end{bmatrix} = (A + GC)\begin{bmatrix} e_i \\ e_\psi \end{bmatrix} + \Delta\omega_r \begin{bmatrix} 0 & -\dfrac{J}{\varepsilon} \\ 0 & J \end{bmatrix}\begin{bmatrix} \hat{i}_s \\ \hat{\psi}_r \end{bmatrix}, \tag{4}$$

where $e_i = i_s - \hat{i}_s$, $e_\psi = \psi_r - \hat{\psi}_r$, and $\varepsilon = \delta L_s L_r / L_m$.

The speed adaptive law [18] can be obtained from the state error equation (4) by using Lyapunov stability theorem:

$$\begin{cases} \hat{\omega}_r = K_p(\varepsilon_1 - \varepsilon_2) + K_i \displaystyle\int (\varepsilon_1 - \varepsilon_2)\mathrm{d}t, \\ \varepsilon_1 = \left(e_{i_{s\alpha}}\hat{\psi}_{r\beta} - e_{i_{s\beta}}\hat{\psi}_{r\alpha}\right), \\ \varepsilon_2 = \left(e_{\psi_{s\alpha}}\hat{\psi}_{r\beta} - e_{\psi_{s\beta}}\hat{\psi}_{r\alpha}\right). \end{cases} \tag{5}$$

Note that it is impossible to obtain actual rotor-flux linkage, if it is assumed that the estimated flux linkage is equal to the actual flux linkage, $\varepsilon_2 = 0$, and the traditional speed adaptive law is obtained:

$$\hat{\omega}_r = K_p \varepsilon_1 + K_i \int \varepsilon_1 \mathrm{d}t. \tag{6}$$

When the motor operates at the medium-high speed, the flux linkage error term is small, which has little impact on the estimation of flux linkage when it is ignored. However,

when the motor operates at low speed, the rotor-flux linkage error will increase significantly, which leads to inaccurate observation.

3. Design of Speed Adaptive Law

3.1. Observer Based on Traditional Adaptive Law. By applying Laplace transform in the state error equation (4), we obtain

$$s\begin{bmatrix} e_i \\ e_\psi \end{bmatrix} = (A + \mathrm{GC})\begin{bmatrix} e_i \\ e_\psi \end{bmatrix} + \Delta\omega_r J\widehat{\psi}_r \begin{bmatrix} -\dfrac{J}{\varepsilon} \\ J \end{bmatrix}, \tag{7}$$

where s is the differential divisor.

A closed-loop system composed of the error equation and speed adaptive link can be established by formulas (6) and (7). The system structure of this system is shown in Figure 1.

As shown in Figure 1, the input of the transfer function of the linear time-invariant forward path is $\Delta\omega_r J\widehat{\psi}_r$. The output is the stator current error e_i, and the formula below is obtained:

$$G(s) = \frac{e_i}{\Delta\omega_r J\widehat{\psi}_r}. \tag{8}$$

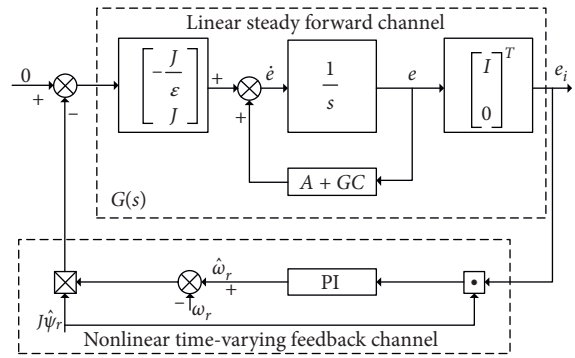


Figure 1: Traditional observer structure in the stationary coordinate system.

By expanding formula (7) in s domain, the following formula is obtained:

$$\begin{cases} sIe_i = (A_{11} + G_1)e_i + A_{12}e_\psi - \dfrac{\Delta\omega_r}{\varepsilon} J\widehat{\psi}_r, \\ sIe_\psi = (A_{11} + G_1)e_i + A_{12}e_\psi - \Delta\omega_r J\widehat{\psi}_r. \end{cases} \tag{9}$$

Specific expression of transfer function of the linear time-invariant forward path is obtained by eliminating e_ψ in the simultaneous equations (9):

$$G(s) = \frac{s}{\varepsilon}\left[s^2 I - s(A_{11} + G_1 + A_{22}) + A_{22}\left(A_{11} + G_1 + \frac{A_{21} + G_2}{\varepsilon}\right)\right]^{-1}. \tag{10}$$

To facilitate the analysis of the stability of the full-order observer, the state error formula (7) is transformed into the rotor-flux linkage-oriented synchronously rotating coordinate system:

$$s\begin{bmatrix} e_i \\ e_\psi \end{bmatrix} = (A' + \mathrm{GC})\begin{bmatrix} e_i \\ e_\psi \end{bmatrix} + \Delta A'\begin{bmatrix} \widehat{i}_s \\ \widehat{\psi}_r \end{bmatrix}, \tag{11}$$

where $A' = \begin{bmatrix} A_{11} - \omega_1 J & A_{12} \\ A_{21} & A_{22} - \omega_1 J \end{bmatrix}$ and $\Delta A' = A' - \widehat{A}'$. The state variables are the components under synchronously rotating coordinate systems m and t.

If the transfer function of the forward path is expressed by $G'(s)$ in coordinate systems m and t, formula (8) can be transformed into the following [10]:

$$\begin{bmatrix} e_{\mathrm{ism}} \\ e_{\mathrm{ist}} \end{bmatrix} = \begin{pmatrix} G'_{11}(s) & G'_{12}(s) \\ G'_{21}(s) & G'_{22}(s) \end{pmatrix}\begin{bmatrix} 0 \\ \widehat{\psi}_r \end{bmatrix}\Delta\omega_r. \tag{12}$$

The elements of the transfer function $G'(s)$ matrix can be obtained by error equations under synchronously rotating coordinate systems [11].

The transfer function from m-axis component of stator current error to speed difference is expressed by $G'_m(s)$. The transfer function from t-axis component of stator current error to speed difference is expressed by $G'_t(s)$:

$$\begin{cases} G'_m(s) = \dfrac{e_{\mathrm{ism}}}{\Delta\omega_r} = G'_{12}(s)\widehat{\psi}_r, \\ G'_t(s) = \dfrac{e_{\mathrm{ist}}}{\Delta\omega_r} = G'_{22}(s)\widehat{\psi}_r. \end{cases} \tag{13}$$

The adaptive law equation is obtained by transforming the traditional adaptive law into coordinate systems m and t by coordinate transformation, as shown in the following equation:

$$\widehat{\omega}_r = -\left(K_p + K_i \int \mathrm{d}t\right)\left(i_{\mathrm{st}} - \widehat{i}_{\mathrm{st}}\right)\psi_{\mathrm{rm}}. \tag{14}$$

The structure diagram of the traditional full-order observer in the synchronously rotating coordinate system can be obtained by synthesizing equations (13) and (14), as shown in Figure 2.

3.2. Design of Improved Speed Adaptive Law. It can be seen from Figure 2 that the traditional adaptive full-order observer is a closed-loop system with single input and single output. In the closed-loop system, only the torque current

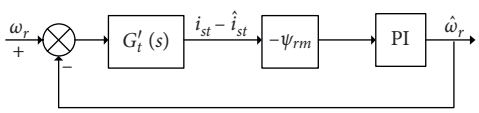


Figure 2: System structure diagram of the traditional full-order observer.

error component is involved in speed identification, and excitation current error component is not a part of the speed identification system.

By introducing the excitation current error component into the traditional speed identification system, equation (14) can be modified as follows:

$$\widehat{\omega}_r = -\left(K_p + K_i \int \mathrm{d}t\right)\left[\left(i_{\mathrm{st}} - \widehat{i}_{\mathrm{st}}\right)\psi_{\mathrm{rm}} + M\left(i_{\mathrm{sm}} - \widehat{i}_{\mathrm{sm}}\right)\right]. \tag{15}$$

If $M = L_r\psi_{\mathrm{st}}$ and the introduced compensation term $M(i_{\mathrm{sm}} - \widehat{i}_{\mathrm{sm}})$ is transformed into the static coordinate system, the compensation term is approximately equal to ε_2[18]. So, it is the negligence of the flux linkage error term in the adaptive law of the traditional speed identification system that leads to the lack of excitation current error component in the synchronously rotating coordinate system, resulting in the inaccurate low-speed observation.

Considering that the actual value of rotor-flux linkage cannot be measured in actual application, the rotor-flux linkage error term ε_2 in the static coordinate system is transformed into detectable stator current:

$$\begin{aligned}\varepsilon_2 &= \left(e_{\psi_{s\alpha}}\widehat{\psi}_{r\beta} - e_{\psi_{s\beta}}\widehat{\psi}_{r\alpha}\right) = \psi_{r\alpha}\widehat{\psi}_{r\beta} - \widehat{\psi}_{r\alpha}\psi_{r\beta} \\ &= \frac{\psi_{r\alpha}\widehat{\psi}_{r\beta} - \widehat{\psi}_{r\alpha}\psi_{r\beta}}{|\psi_r| \cdot |\widehat{\psi}_r|}|\psi_r||\widehat{\psi}_r| \\ &= (\cos\theta\sin\widehat{\theta} - \sin\theta\cos\widehat{\theta})|\psi_r||\widehat{\psi}_r| = \sin\Delta\theta|\psi_r||\widehat{\psi}_r|,\end{aligned} \tag{16}$$

where $|\psi_r|$ is the rotor-flux linkage vector module value and $\Delta\theta$ is the difference between the observed rotor-flux linkage vector angle $\widehat{\theta}$ and the actual rotor-flux linkage vector angle θ.

The rotor-flux linkage vector angle difference can be replaced by the stator current vector angle difference [19]:

$$\sin\Delta\theta = \frac{i_{s\alpha}\widehat{i}_{s\beta} - \widehat{i}_{s\alpha}i_{s\beta}}{|i_s||\widehat{i}_s|}. \tag{17}$$

By introducing equation (17) into equation (16), the following equation is obtained:

$$\varepsilon_2 = \frac{i_{s\alpha}\widehat{i}_{s\beta} - \widehat{i}_{s\alpha}i_{s\beta}}{|i_s||\widehat{i}_s|}|\psi_r||\widehat{\psi}_r| = H\left(i_{s\alpha}\widehat{i}_{s\beta} - \widehat{i}_{s\alpha}i_{s\beta}\right), \tag{18}$$

where $H$is the weight coefficient. The accuracy and dynamic performance of the observer can be improved by adjusting the H value [20]. The typical value of parameter h can be designed as shown in the following equation:

$$\begin{cases} 0 < H < 0.05, & (\omega_1 < 0), \\ H = 0, & (\omega_1 \geq 0). \end{cases} \tag{19}$$

4. Stability Analysis and Improvement of Observer

4.1. Analysis of the Unstable Range for Full-Order Observer. In theory, the stability of the full-order observer can be improved by weighting and compensating the adaptive law. However, the commissioning cycle will be extended, and there is a great blindness if the weight coefficient is adjusted in real time based on open-loop observation ($G = 0$). In addition, to improve the convergence speed of full-order observer speed identification, the open-loop gain is usually set to a large value. Considering that the root locus of the closed-loop transfer function starts from the open-loop pole and eventually tends to the open-loop zero point, an unreasonable weight coefficient will lead to a positive real part of the open-loop zero point of the observer, which causes instability as the closed-loop root locus of the full-order observer tends to open-loop zero point due to the large open-loop gain.

To analyze the unstable region of the open-loop observer and reasonably configure the feedback gain matrix to form a closed-loop full-order observer to eliminate the low-speed unstable region, the transfer function (10) of the linear time-invariant forward path can be simplified as follows:

$$G(s) = \frac{s}{\varepsilon}\left[s^2 I - s(aI + bI) + cI + dI\right]^{-1}, \tag{20}$$

where

$$\begin{aligned} a &= -g_1 + \frac{R_r}{\delta L_r} + \frac{R_s}{\delta L_s}, \\ b &= -g_2 - \omega_r, \\ c &= -\frac{R_r}{L_r}\left(-\frac{R_s}{\delta L_s} + g_1 + \frac{g_3}{\varepsilon}\right) - \omega_r\left(g_2 + \frac{g_4}{\varepsilon}\right), \\ d &= \omega_r\left(-\frac{R_s}{\delta L_s} + g_1 + \frac{g_3}{\varepsilon}\right) - \frac{R_r}{L_r}\left(g_2 + \frac{g_4}{\varepsilon}\right). \end{aligned} \tag{21}$$

According to Popov's hyperstability theorem, to ensure the asymptotic stability of the speed identification system, the transfer function of the linear time-invariant forward path should be a strictly positive real function:

$$G(j\omega) + G^*(j\omega) > 0, \quad \forall\omega > 0. \tag{22}$$

By introducing equation (20) and $s = j\omega_1$ into equation (20), a simplified equation is obtained:

$$\begin{cases} a > 0, \\ \omega_1^2 > \left(-\dfrac{d}{a}\right)^2, \end{cases} \tag{23}$$

where ω_1 is the synchronous angular frequency, so $\omega_c = -d/a$ is the critical angular frequency.

Formula (23) is the stability condition of the speed identification system, and the constraint condition $a > 0$ is naturally satisfied under open-loop observation ($G = 0$). If the motor operates in the forward rotation state and the synchronous frequency is positive, the unstable region of the open-loop observation speed identification system is as follows:

$$0 < \omega_1 < \omega_c = \frac{(R_s/\delta L_s)}{(R_s/\delta L_s) + (R_r/\delta L_r)}\omega_r < \omega_r. \tag{24}$$

The relationship between the electromagnetic torque and the speed of the induction motor is presented as follows:

$$T_e = n_p \frac{\psi_r^2}{R_r}(\omega_1 - \omega_r). \tag{25}$$

By introducing the boundary condition of the unstable region into equation (25), it is obtained that

$$\begin{cases} T_e = -n_p \dfrac{\psi_r^2}{R_r}\omega_r, \\ T_e = -n_p \dfrac{R_r/\delta L_r}{R_r/\delta L_r + R_s/\delta L_s}\dfrac{\psi_r^2}{R_r}\omega_r. \end{cases} \tag{26}$$

The graph of the unstable region is plotted with electromagnetic torque and speed, as shown in Figure 3(a). The shaded part in the figure is the unstable region, and the expression of the boundary line is shown in expression (26). In this case, the actual speed is greater than the synchronous speed and the slip frequency is negative, which means that the motor is in the dynamic braking state (unstable state).

4.2. Stability Improvement of Full-Order Observer. From the stability constraint expression (23), the stability of the full-order observer is subjected to the design of the feedback gain matrix. The stability of the observer can be improved by configuring a feedback gain matrix. To meet the low-speed stability requirements of the motor operation, the critical angular frequency $\omega_c$is set to zero. At this point, the two boundary lines in Figure 3(a) coincide and the unstable region disappears, as shown in Figure 3(b). The stability constraint can be simplified as follows:

$$\begin{cases} g_1 < \dfrac{R_s}{\delta L_s} + \dfrac{R_r}{\delta L_r}, \\ \dfrac{R_r}{L_r}\left(g_2 + \dfrac{g_4}{\varepsilon}\right) = \omega_r\left(-\dfrac{R_s}{\delta L_s} + g_1 + \dfrac{g_3}{\varepsilon}\right). \end{cases} \tag{27}$$

According to this principle, the elements of the feedback gain matrix can be configured as follows [11]:

$$\begin{cases} g_1 = \dfrac{R_s L_r^2 + R_r L_m^2}{\delta L_s L_r^2} - k\dfrac{R_r}{L_r}, \\ g_2 = -k\omega_r, \\ g_3 = -\dfrac{L_m R_r}{L_r}, \\ g_4 = 0. \end{cases} \tag{28}$$

where k is the ratio of the observer pole to motor pole.

According to this design scheme, although global stability is achieved, the observer pole position is moved to the position close to the origin, which reduces the convergence speed of the system.

It can be seen from expression (28) that since the feedback gain matrix itself is time-varying and constantly updated, complicated element design will inevitably reduce its convergence performance. Therefore, in this paper, the feedback gain matrix is simplified.

$$\begin{cases} g_1 = k\dfrac{R_s L_r^2 + R_r L_m^2}{\delta L_s L_r^2}, \\ g_2 = 0, \\ g_3 = -\dfrac{L_m R_r}{L_r}, \\ g_4 = 0. \end{cases} \tag{29}$$

The final design scheme of the adaptive full-order observer can be obtained by synthesizing expressions (5), (18), and (29), as shown in Figure 4. The design scheme not only solves the problem of low-speed instability by reasonably designing the gain matrix but also improves the dynamic performance of the system by combining with the improved weighted adaptive law.

5. System Simulations and Experiments

5.1. System Simulations. In this paper, simulation of the decoupling vector control system of the full-order observer-based induction motor is carried out, and the simulation model of the control algorithm is constructed using MATLAB/SIMULINK, as shown in Figure 5.

In the simulation model, basic parameters of the induction motor are set as follows: $u_N = 380$ V, $P_N = 3{*}746$ W, $f = 50$ Hz, $R_s = 0.435\,\Omega$, $R_r = 0.816\,\Omega$, $L_m = 0.069$ H, $L_{lr} = L_{ls} = 0.002$ H, $J = 0.01$ kg·m^2, and $p = 2$.

Figures 6(a) and 6(b) are the speed waveforms of the traditional full-order observer control strategy and the

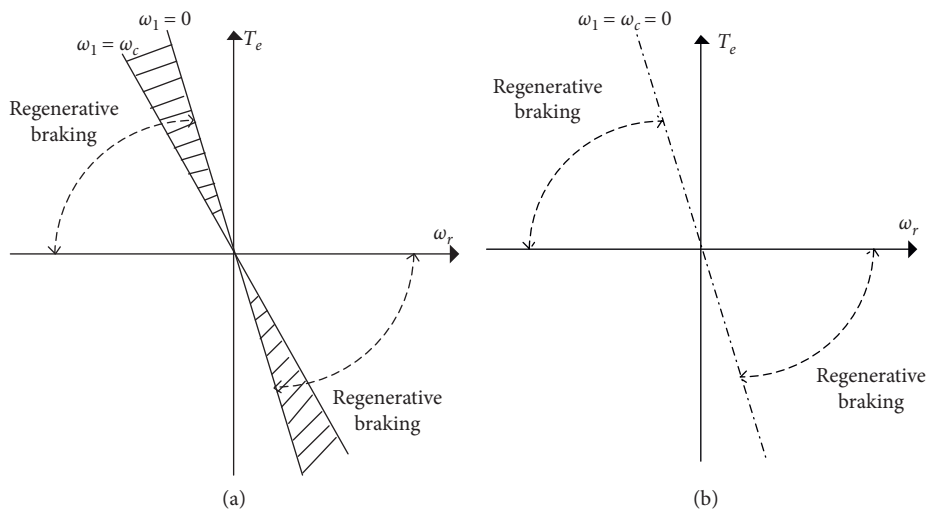


Figure 3: Diagram of the asynchronous motor $\omega_r - T_e$ based on(a) open-loop speed observer and (b) closed-loop speed observer.

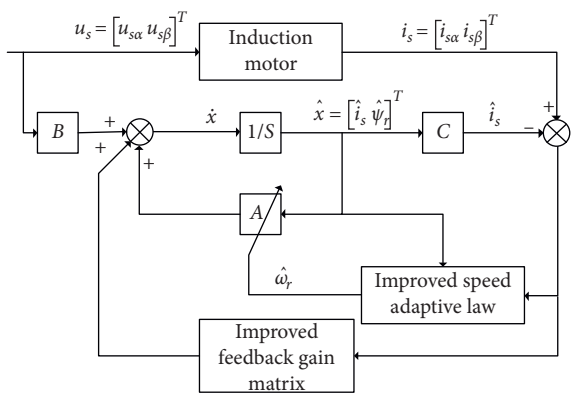


Figure 4: The improved system structure diagram of the adaptive full-order observer.

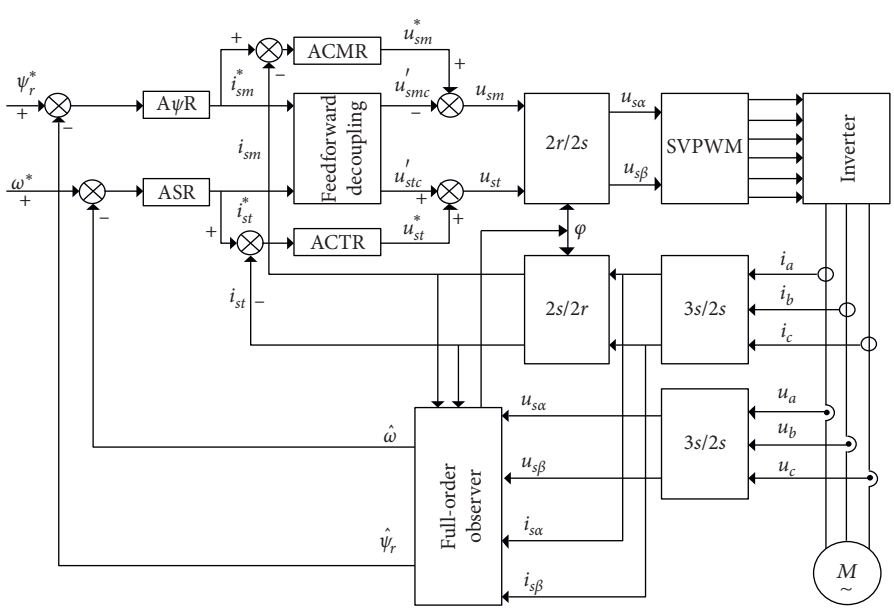


Figure 5: Control block diagram of the system.

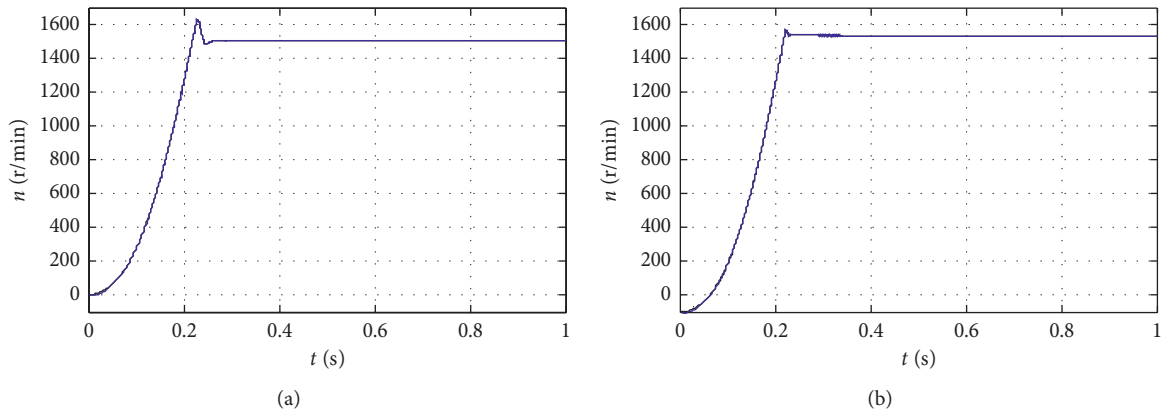


Figure 6: Speed waveform diagram of the control system at high speed. (a) Speed waveform of the traditional observer. (b) Improved full-order observer speed waveform.

improved full-order observer control strategy at high speed. From the speed graphs of two control strategies, in the high-speed and no-load state, the motor speed rises steadily to 1500 r/min in 0.25 s, and the overshoot of the improved full-order observer is lower than that of the traditional observer. At this point, the actual speed curve and the estimated speed curve of the two control strategies basically coincide, and both speed identification systems can accurately track the real speed.

In the low-speed regenerative braking mode, the given speed is set to 100r/min and the given flux linkage to 0.9 Wb. From formula (25) we know that the critical value of power-generating load is −27 N·m. As a result, the load applied to the motor is set to −30 N·m.

Figure 7 is the speed waveform of the control system in the low-speed regenerative mode. To verify the stability of the control system under the regenerative state, the power-generating load is used for simulation experiment. As shown in the figure, the motor starts with no load, and then the speed is maintained at 100 r/min. At 0.5 s, the power-generating load of −30 N·m is suddenly applied to the motor. As the load applied exceeds the critical value, the traditional observer enters the unstable region. The observed speed becomes divergent and no longer converges to the actual speed, while the improved full-order observer converges to the actual speed stably. This is consistent with previous theoretical analysis, proving that the improved full-order observer control system has good low-speed stability.

Figure 8 shows the component diagram of rotor-flux linkage of the control system in the low-speed regenerative mode. When the power-generating load is suddenly applied at 0.5 s, the flux linkage of the traditional observer diverges, while the flux linkage of the improved full-order observer has accurate estimation without DC bias and error in integral initial value of open-loop estimation.

Figures 9 and 10 are the speed waveforms and their partial enlarged drawings of the improved full-order observer when load is added or reduced at low speed. In the low-speed state, the motor starts at no load and then steadily rises to 100 r/min at low speed. At 0.4 s, the load torque of the motor steps from 0 to −30 N·m; at 0.6 s, the load torque steps from −30 N·m to 30 N·m. In this process, the estimated speed still tracks the actual speed in real time, showing that the control system has good dynamic performance when the load is added or reduced.

Figure 11(a) is the speed switch waveform at low speed. At 0.4 s, the given speed of the control system is stepped from 50 r/min to 30 r/min and from 30 r/min to 10 r/min at 0.6 s. In the figure, the control system not only can operate stably at extremely low speed but also has fast speed and small overshoot in the switching process. As shown in Figure 11(b), after the flux linkage is stabilized, the influence speed change is neglectable. It can be seen that the improved control scheme not only improves the dynamic performance but also has good low-speed stability.

5.2. System Experiments. The improved control algorithm is tested on a 5 kW induction motor doubly-fed platform, as shown in Figure 12. Motor 1 is the test motor, and Motor 2 the load motor. Some parameters of the motors in the experiment are as follows: $u_N = 380\,\text{V}$, $P_N = 5\,\text{kW}$, $f = 50\,\text{Hz}$, $I_N = 11.1\,\text{A}$, $p = 2$, $n_N = 1440\,\text{r/min}$. In the test, Motor 1 works in the speed identification state and uses the speed obtained from speed identification to conduct closed-loop vector control. The stability of the control system at low speed is verified by observing the actual speed and estimated speed of Motor 1.

Figure 13 shows the three-phase stator current waveform of the induction motor at a low speed of 100 r/min. The three-phase stator current waveform at low speed is symmetrical and basically stable.

Figure 14 shows the waveform of the actual speed. When the given speed is switched from 600 r/min to 200 r/min and 100 r/min, respectively, the dynamic performance of the system is good during the whole process, and the motor operation is still stable when switched to the low-speed mode, which proves the effectiveness of the improved control strategy.

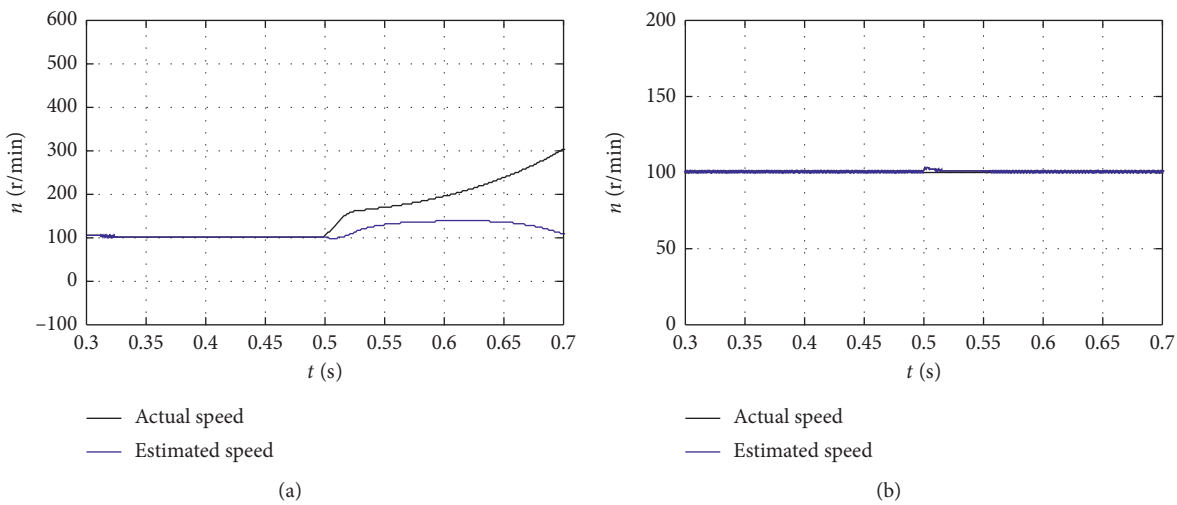


Figure 7: Speed waveform diagram of the control system in the low-speed regeneration mode. (a) Speed waveform of the traditional observer. (b) Improved full-order observer speed waveform.

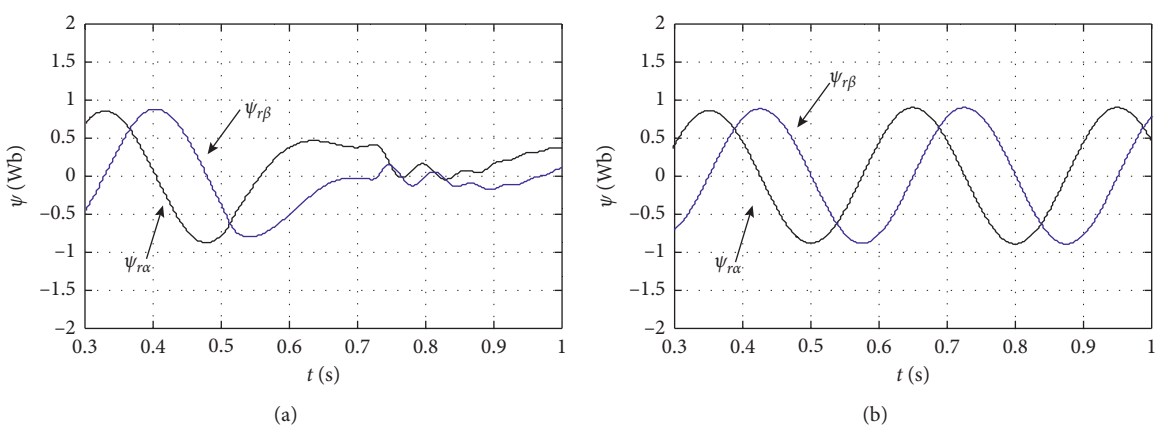


Figure 8: Rotor-flux component diagram under the low-speed regeneration mode of the control system. (a) Rotor-flux linkage diagram of the traditional observer. (b) Improved rotor-flux linkage diagram of the full-order observer.

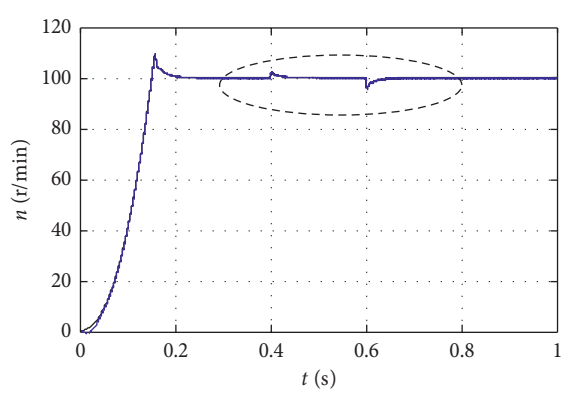


Figure 9: Waveform diagram of load speed at low speed after Improvement.

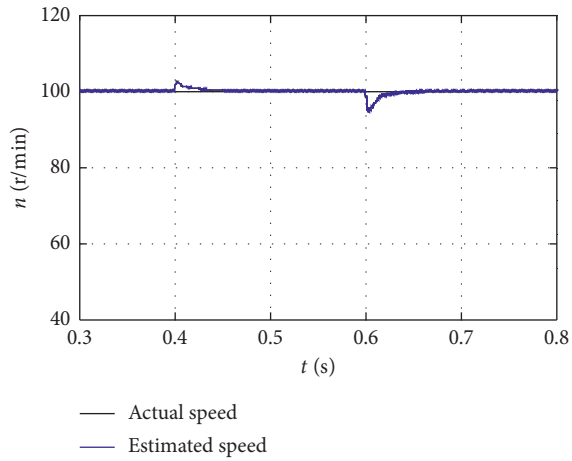


Figure 10: Local enlargement diagram of load addition and subtraction at improved low speed.

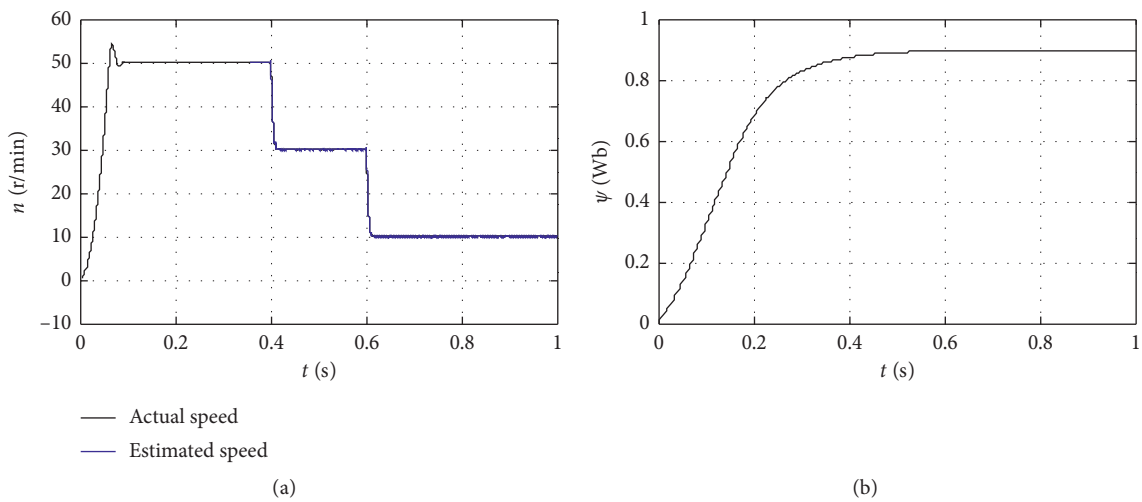


Figure 11: The improved waveform of speed switch and flux at low speed. (a) Speed switching waveform at low speed. (b) Flux linkage diagram under speed switching.

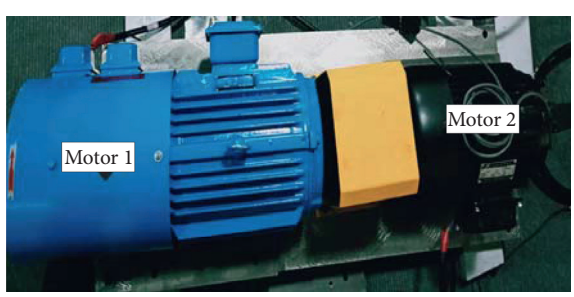


Figure 12: Asynchronous motor test platform.

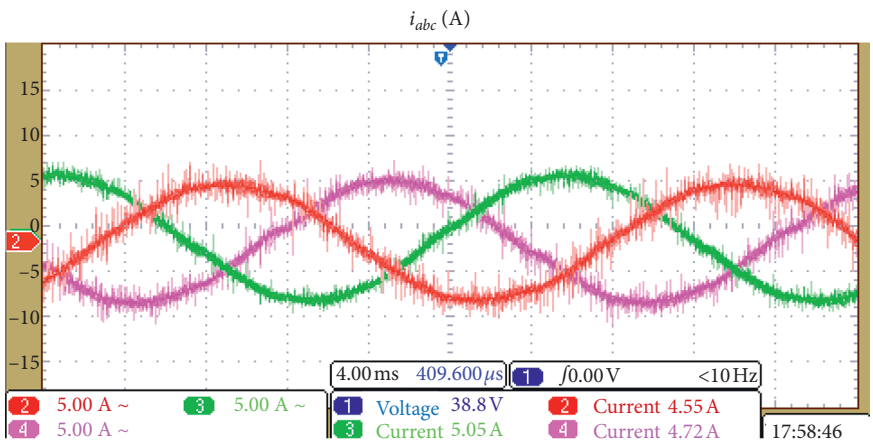


Figure 13: Three-phase stator current waveform of the asynchronous motor at low speed.

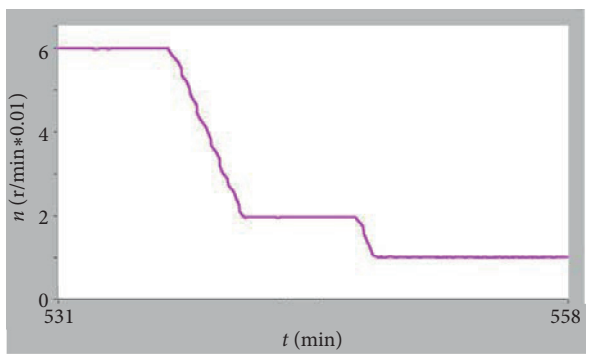


Figure 14: Waveform of actual speed and speed.

6. Conclusion

In this paper, a control strategy for low-speed stability optimization of the induction motor based on the full-order observer is proposed. The low-speed instability of the full-order observer in the speed identification system is analyzed. The feedback gain matrix is designed to eliminate the unstable region of the control system, and the feedback gain matrix is simplified to improve the convergence speed. Combined with the weighted adaptive law, the good dynamic and static performance of the control system is achieved. The simulation results show that the control strategy can improve the stability at low speed and increase the accuracy of speed identification.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. Zhang, H. Zhang, and Z. Li, *High Performance Control Technology of Asynchronous Motor without Speed Sensor*, Mechanical Industry Press, Beijing, China, 2015.

[2] X. Zou, P. Zhu, and Y. Kang, "Speed sensorless vector control of induction motor based on voltage decoupling principle," *Proceedings of the CSEE*, vol. 25, no. 14, pp. 99-100, 2005.

[3] Y. B. Zbede, S. M. Gadoue, and D. J. Atkinson, "Model predictive MRAS estimator for sensorless induction motor drives," *IEEE Transactions on Industrial Electronics*, vol. 26, no. 1, pp. 3511–3521, 2016.

[4] H. Zhang, W. Liu, and J. Peng, "Saturation salient response analysis and rotor position estimation of surface mount permanent magnet synchronous motor based on square wave voltage signal injection," *Transactions of China Electrotechnical Society*, vol. 32, no. 16, pp. 106–114, 2017.

[5] H. Kubota, K. Matsuse, and T. Nakano, "DSP-based speed adaptive flux observer of induction motor," *IEEE Transactions on Industry Applications*, vol. 29, no. 2, pp. 344–348, 1993.

[6] H. Tajima, G. Guidi, and H. Umida, "Consideration about problems and solutions of speed estimation method and parameter tuning for speed sensor-less vector control of induction motor drives," *IEEE Transactions on Industry Applications*, vol. 38, no. 2, pp. 1282–1289, 2002.

[7] X. Li, S. Yang, and P. Cao, "Stability analysis and design of adaptive observer for asynchronous drive speed at low speed,"

Transactions of China Electrotechnical Society, vol. 33, no. 23, pp. 5391–5400, 2018.

[8] W. Song, G. Yao, and W. Zhou, "Pole configuration method of asynchronous motor full-order state observer," *Electric Machines and Control Application*, vol. 35, no. 9, pp. 06–10, 2008.

[9] J. Maes and J. A. Melkebeek, "Speed sensor-less direct torque control of induction motors using an adaptive flux observer," *IEEE Transactions on Industry Applications*, vol. 36, no. 4, pp. 778–785, 2000.

[10] W. Song, J. Zhou, and H. Zhu, "Stability of induction motor low-speed generating operation based on adaptive full-order observer," *Transactions of China Electrotechnical Society*, vol. 29, no. 3, pp. 198–205, 2014.

[11] S. Suwankawin and S. Sangwongwanich, "Design strategy of an adaptive full-order observer for speed-sensorless induction-motor Drives-tracking performance and stabilization," *IEEE Transactions on Industrial Electronics*, vol. 53, no. 1, pp. 96–119, 2006.

[12] Y. Zhang, Z. Yin, G. Li et al., "A novel speed estimation method of induction motors using real-time adaptive extended Kalman filter," *Journal of Electrical Engineering & Technology*, vol. 13, no. 1, pp. 287–297, 2018.

[13] J. Chen and J. Huang, "Online decoupled stator and rotor resistances adaptation for speed sensorless induction motor drives by a time-division approach," *IEEE Transactions on Power Electronics*, vol. 32, no. 6, pp. 4587–4599, 2017.

[14] H. Wang, X. Ge, and Y.-C. Liu, "Second-order sliding-mode MRAS observer-based sensorless vector control of linear induction motor drives for medium-low speed maglev applications," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 12, pp. 9938–9952, 2018.

[15] M. Norambuena, J. Rodriguez, Z. Zhang, F. Wang, C. Garcia, and R. Kennel, "A very simple strategy for high-quality performance of AC machines using model predictive control," *IEEE Transactions on Power Electronics*, vol. 34, no. 1, pp. 794–800, 2019.

[16] S. A. Davari, D. A. Khaburi, F. Wang, and R. M. Kennel, "Using full order and reduced order observers for Robust sensorless predictive torque control of induction motors," *IEEE Transactions on Power Electronics*, vol. 27, no. 7, pp. 3424–3433, 2012.

[17] H. Yang, Y. Zhang, P. D. Walker, N. Zhang, and B. Xia, "A method to start rotating induction motor based on speed sensorless model-predictive control," *IEEE Transactions on Energy Conversion*, vol. 32, no. 1, pp. 359–368, 2017.

[18] W. Chen, Y. Yu, and R. Yang, "Research on low speed stability of adaptive full-order observer algorithm for asynchronous motor," *Proceedings of the CSEE*, vol. 30, no. 36, pp. 33–40, 2010.

[19] W. Sun, Y. Yu, G. Wang, B. Li, and D. Xu, "Design method of adaptive full order observer with or without estimated flux error in speed estimation algorithm," *IEEE Transactions on Power Electronics*, vol. 31, no. 3, pp. 2609–2626, 2016.

[20] W. Sun, J. Gao, and Y. Yu, "Robustness improvement of speed estimation in speed sensorless induction motor drives," *IEEE Transactions on IndustryApplications*, vol. 52, no. 3, pp. 2525–2536, 2015.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8849395, 14 pages
https://doi.org/10.1155/2020/8849395

Research Article

Stability of 1-Bit Compressed Sensing in Sparse Data Reconstruction

Yuefang Lian,[1] Jinchuan Zhou,[1] Jingyong Tang,[2] and Zhongfeng Sun[3]

[1]*Department of Statistics, School of Mathematics and Statistics, Shandong University of Technology, Zibo 255000, China*
[2]*School of Mathematics and Statistics, Xinyang Normal University, Xinyang 464000, China*
[3]*Department of Information and Computing Science, School of Mathematics and Statistics, Shandong University of Technology, Zibo 255000, China*

Correspondence should be addressed to Jinchuan Zhou; jinchuanzhou@163.com

Received 25 September 2020; Revised 15 October 2020; Accepted 6 November 2020; Published 25 November 2020

Academic Editor: Guoqiang Wang

1-bit compressing sensing (CS) is an important class of sparse optimization problems. This paper focuses on the stability theory for 1-bit CS with quadratic constraint. The model is rebuilt by reformulating sign measurements by linear equality and inequality constraints, and the quadratic constraint with noise is approximated by polytopes to any level of accuracy. A new concept called restricted weak RSP of a transposed sensing matrix with respect to the measurement vector is introduced. Our results show that this concept is a sufficient and necessary condition for the stability of 1-bit CS without noise and is a sufficient condition if the noise is available.

1. Introduction

The standard noiseless compressing sensing (CS) model is to solve the following optimization problem:

$$\begin{aligned} &\min \|x\|_0, \\ &\text{s.t. } Ax = y, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ is a sensing (or measurement) matrix and x is a sparse signal requiring robust reconstruction from a given nonadaptive measurement vector y [1–4]. The l_0-minimization problem is well known to be NP-hard. Hence, to overcome this difficulty, a typical treatment is resorting to use l_1-norm. Along this approach, a great deal of algorithms is available, e.g., orthogonal matching pursuit algorithm [5], basis pursuit algorithm [6], iterative hard threshold algorithm [7], and iteratively reweighted least squares algorithm [8]. Moreover, some added assumptions have to be added on the measurement matrix A to ensure that a sparse solution/signal could be exactly recovered by l_1 minimization. These conditions include restricted isometry property [9–11], coherence condition [12], null space property [8, 13, 14], and range space property [15, 16]. In recent research, some work has been done concerning the robust reconstruction condition (RRC) based on the above traditional properties and their variants, e.g., exact reconstruction condition [17], double null space property [18], and null space property [19].

However, the above CS model cannot be adapted in some practical problems; for example, in brain signal processing and sigma-delta converters, only the sign or support of a signal is measured. This motivates one to consider sparse signal recovery through low bits of measurements. An extreme quantization is only one bit per measurement. It gives rise to the theory of 1-bit compressed sensing (see Boufounos and Baraniuk [20]).

In this paper, we further consider a constrained 1-bit compressed sensing model involved by a noisy constraint. Precisely, let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{l\times n}$ be two given full-row rank matrices. Pick $y \in \{1,-1,0\}^m$ with, $b \in \mathbb{R}^l$ is a given vector, and ε is a positive number. The constrained 1-bit compressed sensing model is described as follows:

$$\begin{aligned} &(P)\min \|x\|_0 \\ &\text{s.t. } \operatorname{sign}(Ax) = y, \\ &\|b - Bx\|_2 \le \varepsilon, \end{aligned} \tag{2}$$

where the last term $\|b - Bx\|_2 \le \varepsilon$ stands for a noisy constraint. The corresponding convex relaxed problem via l_1-norm is expressed as

$$\begin{aligned} &\min \|x\|_1, \\ &\text{s.t. } \operatorname{sign}(Ax) = y, \\ &\|b - Bx\|_2 \le \varepsilon. \end{aligned} \tag{3}$$

Compared with the recovery of a given signal, it is equally important to study whether the recovered signal is stable. The stability of recovery means that recovery errors stay under control even if the measurements are slightly inaccurate and the data are not exactly sparse. Recent stability study for CS can be found in [21–25]. However, few theoretical results are available on the stability of 1-bit CS. In general, it is impossible to exactly reconstruct a sparse signal by only using 1-bit information. For example, if $\operatorname{sign}(Ax^*) = (1,1)$, then any sufficiently small perturbation $x^* + v$ is also positive and hence satisfies the requirement. Hence, we turn our attention to recover part of the information in 1-bit CS, such as support set or sign of a target signal. Due to this reason, the following criterion,

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \le \Delta, \tag{4}$$

where $x \ne 0$ and $x^* \ne 0$ and Δ denotes a sufficient small positive scalar and has been widely used in the 1-bit CS literature. Inspired by this observation, the problem (1) is said to be stable for noisy reconstruction, if for any nonzero vector $x \in \mathfrak{R}^n$, there is a nonzero solution x^* of (3) such that

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \le \tau(x)\left(C_1\sigma_k(x)_1 + C_2\varepsilon\right), \tag{5}$$

where C_1 and C_2 are constant depending on the primal problem data $(A, y, \varepsilon, B, b)$. If $\varepsilon = 0$ and x is k-sparse, then the right side of (5) is zero and hence $x/\|x\|_2 = x^*/\|x^*\|_2$, which in turn implies that $\operatorname{sign}(x) = \operatorname{sign}(x^*)$; i.e., the sign of target signals can be exact recovery.

The main target of this paper is to study the necessary and/or sufficient condition for (5). First, a new definition called restricted weak RSP with respect to y is introduced. Our results show that, for 1-bit CS, this condition is sufficient and necessary condition for stability if there is no noise, while it is sufficient if the noise is available. The analysis is based on the duality theory of linear programming and the fact that the ball constraint can be approximated by polytopes to any level of accuracy.

The notations used in this paper are standard. Let $\mathbb{R}^n_+$ be the set of nonnegative vectors in $\mathbb{R}^n$. Given a set S, $|S|$ denotes the cardinality of S. The l_0-norm $\|x\|_0$ counts the number of nonzero components of x, and the l_1-norm of x is defined as $\|x\|_1 := \sum_{i=1}^n |x_i|$. Let e stand for a vector of ones, i.e., $e = (1,\dots,1)^T$. For a vector x, write $x^+ := \max\{x, 0\}$ and $x^- := \max\{-x, 0\}$. For any two norms $\|\cdot\|_p$ and $\|\cdot\|_q$ with $p, q \ge 1$, the induced matrix norm $\|A\|_{p\longrightarrow q}$ is defined as $\|A\|_{p\longrightarrow q} := \max_{\|x\|_p \le 1}\|Ax\|_q$. A convex combination between the points x_1 and x_2 is written as $[x_1, x_2]$, i.e.,

$$[x_1, x_2] := \left\{\lambda x_1 + (1-\lambda)x_2 \,\middle|\, \lambda \in [0,1]\right\}. \tag{6}$$

Given a vector $y = \{1,-1,0\}^m$, let

$$\begin{aligned} J_+(y) &:= \{i\colon y_i = 1\}, \\ J_-(y) &:= \{i\colon y_i = -1\}, \\ J_0(y) &:= \{i\colon y_i = 0\}. \end{aligned} \tag{7}$$

The sign function is defined as

$$\operatorname{sign}(t) := \begin{cases} 1, & t > 0, \\ -1, & t < 0, \\ 0, & t = 0, \end{cases} \tag{8}$$

and $\operatorname{sign}(x)_i := \operatorname{sign}(x_i)$ where $x \in \mathfrak{R}^n$ and $i = 1,\dots,n$. The projection of x onto a convex set S is denoted by $\pi_S(x)$, i.e., $\pi_S(x) := \operatorname{argmin}_{z\in S}\|x - z\|_2$. Denote by $(S_1 \cup S_2)^c$ the complement of $S_1 \cup S_2$ in $\{1, 2, \dots, n\}$. The error of the best k-term approximation of a vector x is defined as

$$\sigma_k(x)_1 := \inf_u \{\|x - u\|_1 \colon \|u\|_0 \le k\}. \tag{9}$$

The Hausdorff metric of two sets $M_1, M_2 \subseteq \mathfrak{R}^n$ is

$$d^{\mathscr{H}}(M_1, M_2) := \max\left\{ \sup_{x\in M_1} \inf_{z\in M_2} \|x - z\|_2, \sup_{x\in M_2} \inf_{z\in M_1} \|x - z\|_2 \right\}. \tag{10}$$

Robinson's constant is defined as follows:

$$\sigma_{\alpha_1,\alpha_2}(M', M'') := \max_{N\subseteq\{1,\dots,m\}} \mu_{\alpha_1,\alpha_2}\left(\begin{bmatrix} I_N & 0 \\ -I & 0 \end{bmatrix}, \begin{bmatrix} M' \\ M'' \end{bmatrix}^T \right), \tag{11}$$

where

$$\begin{aligned} \mu_{\alpha_1,\alpha_2}(P, Q) &:= \max_{\|(b,d)\|_{\alpha_2} \le 1, (b,d)\in F} \left(\min_{z\in\mathbb{R}^q} \left\{ \|z\|_{\alpha_1} \colon Pz \le b, Qz = d \right\} \right), \\ F &:= \left\{ (b, d) \mid Pz \le b, Qz = d \text{ for some } z \in \mathbb{R}^q \right\}. \end{aligned} \tag{12}$$

2. Reformulation and Approximation of (3)

The 1-bit CS is NP-hard and hence is difficult to solve precisely. It motivates us to reformulate the 1-bit CS problem by removing the sign function. The advantage of such a reformulation is yielding a decoding method based on the theory of linear programming.

Given sign measurements $y \in \{-1, 1, 0\}^m$, denote by A^+, A^-, and A^0 the submatrices of A in which their rows are corresponding to index sets $J_+(y)$, $J_-(y)$, and $J_0(y)$, respectively. For simplification of notations, we simply use J_+, J_0, and J_- to denote $J_+(y)$, $J_0(y)$, and $J_-(y)$, respectively. In the following analysis, we always assume that $J_+ \cup J_- \neq \varnothing$ because otherwise $y = 0$, and nothing is measured in this case.

The constraint sign $(Ax) = y$ can be rewritten equivalently as

$$\begin{aligned} &\operatorname{sign}\left(A^+x\right) = e_{J_+}, \\ &\operatorname{sign}\left(A^-x\right) = -e_{J_-}, \\ &\operatorname{sign}\left(A^0x\right) = 0. \end{aligned} \tag{13}$$

By rearranging the order of the components of y and the order of the associated rows of A if necessary, we may assume without loss of generality that

$$A = \begin{bmatrix} A^+ \\ A^- \\ A^0 \end{bmatrix}, \qquad y = \begin{pmatrix} e_{J_+} \\ -e_{J_-} \\ 0 \end{pmatrix}. \tag{14}$$

It is clear that

$$\left\{x \mid A^+x > 0, A^-x < 0, A^0x = 0\right\} = \bigcup_{\alpha>0}\left\{x \mid A^+x \geq \alpha e, A^-x \leq -\alpha e, A^0x = 0\right\}. \tag{15}$$

In fact, the inclusion "$\supseteq$" is clear. For "$\subseteq$," take x satisfying $A^+x > 0, A^-x < 0, A^0x = 0$. Define

$$\alpha := \min\left\{A_ix, -A_jx \,\middle|\, i \in J_+, j \in J_-\right\}. \tag{16}$$

Clearly, $\alpha > 0$. Thus, $A_ix \geq \alpha$ for all $i \in J_+$ and $-A_jx \geq \alpha$ for all $j \in J_-$; i.e., $A^+x \geq \alpha e$ and $A^-x \leq -\alpha e$. Therefore,

$$x \in \bigcup_{\alpha>0}\left\{x \mid A^+x \geq \alpha e, A^-x \leq -\alpha e, A^0x = 0\right\}. \tag{17}$$

For any fixed $\alpha > 0$, define the following relaxed problem (denoted by α-problem for short),

$$\begin{aligned} &\min \|x\|_1 \\ &\text{s.t. } A^+x \geq \alpha e_{J_+}, A^-x \leq -\alpha e_{J_-}, A^0x = 0, \\ &\|b - Bx\|_2 \leq \varepsilon. \end{aligned} \tag{18}$$

The formula (15) shows that $\mathcal{F} = \cup_{\alpha>0}\mathcal{F}_\alpha$, where $\mathcal{F}$ and $\mathcal{F}_\alpha$ denote the feasible region of the primal problem and the relaxed problem, respectively. In addition, $\mathcal{F}_\beta \subseteq \mathcal{F}_\alpha$ as long as $\beta \geq \alpha$. Thus,

$$\mathcal{F} = \bigcup_{\alpha>0}\mathcal{F}_\alpha \subseteq \lim_{\alpha \longrightarrow 0^+}\mathcal{F}_\alpha = cl\mathcal{F}, \tag{19}$$

where the limit is in the sense of the Painlevé–Kuratowski.

Proposition 1. *A vector x^* is an optimal solution of primal problem (P) if and only if x^* is an optimal solution of β-problem for all $\beta \in (0, \alpha]$, where $\alpha := \min\left\{A_ix^*, -A_jx^* \mid i \in J_+, j \in J_-\right\}$.*

Proof. "$\Rightarrow$." The construction of α ensures that $A^+x^* \geq \alpha e, A^-x^* \leq -\alpha e$, and $A^0x^* = 0$. Hence, for $\forall \beta \leq \alpha$,

$$\begin{aligned} &A^+x^* \geq \alpha e \geq \beta e, \\ &A^-x^* \leq -\alpha e \leq -\beta e, \\ &A^0x^* = 0, \end{aligned} \tag{20}$$

i.e., x^* is a feasible solution of β-problem. Since x^* is an optimal solution of the primal problem, x^* is the optimal solution of β-problem due to $\mathcal{F}_\beta \subset \mathcal{F}$ by (15).

"$\Leftarrow$." Let $\tilde{x}^*$ be an optimal solution of the primal problem. Take $\beta \in (0, \tilde{\alpha})$ where $\tilde{\alpha} := \min\{\alpha, \alpha'\}$ and $\alpha' := \min\left\{A_i\tilde{x}^*, -A_j\tilde{x}^* \mid i \in J_+, j \in J_-\right\}$. Then, $\tilde{x}^*, x^* \in \mathcal{F}_\beta$ due to the monotonicity of $\mathcal{F}_\alpha$ with respect to α. By assumption, x^* is an optimal solution of β-problem. Since $\tilde{x}^* \in \mathcal{F}_\beta$ and is an optimal solution of the primal problem, then x^* is an optimal solution of the primal problem.

Denote by T^* and T^*_α the optimal solution set of (3) and (18), respectively. Following the similar argument as above, we can obtain the following result. □

Corollary 1. *There exists $\alpha > 0$ such that $T^*_\beta \subseteq T^*$ for all $\beta \in (0, \alpha]$.*

The problem (18) by introducing the slack variables r and s can be rewritten equivalently as

$$\begin{aligned} &\min_{x,r,s} \|x\|_1 \\ &\text{s.t. } A^+x \geq \alpha e_{J_+}, A^-x \leq -\alpha e_{J_-}, A^0x = 0, \quad s \leq \varepsilon, r \in s\mathbb{B}, r = b - Bx, s \geq 0, \end{aligned} \tag{21}$$

where $\mathbb{B}$ stands for the unit l_2-ball, i.e., $\mathbb{B} := \{z \in \mathbb{R}^m : \|z\|_2 \le 1\}$. According to the convex set separate theorem, the set $\mathbb{B}$ can be described as an intersection of an infinite number of half spaces, i.e.,

$$\mathbb{B} = \bigcap_{\|a\|_2=1} \left\{z \in \mathbb{R}^m : a^T z \le 1\right\}. \tag{22}$$

Define

$$E_\alpha := \left\{(x,s) : s \le \varepsilon, A^+ x \ge \alpha e_{J_+}, A^- x \le -\alpha e_{J_-}, A^0 x = 0, s \ge 0\right\}. \tag{23}$$

Notice that

$$T_\alpha^* = \left\{x : \|x\|_1 \le \theta_\alpha^*, r \in s\mathbb{B}, r = b - Bx, (x,s) \in E_\alpha\right\}, \tag{24}$$

where θ_α^* denotes the optimal value of (18). Replacing $\mathbb{B}$ in (24) by a polytope $P \supseteq \mathbb{B}$ yields a relaxation of T_α^*, called T_α^P, i.e.,

$$T_\alpha^P := \left\{x : \|x\|_1 \le \theta_\alpha^*, r \in sP, r = b - Bx, (x,s) \in E_\alpha\right\}. \tag{25}$$

The following lemma claims that the polytope T_P can approximate T^* to any level of accuracy, as long as P is chosen suitably.

Lemma 1 (see [25], Corollary 6.5.2). *For any $\varepsilon > 0$, there exists a polytope approximation P of $\mathbb{B}$ satisfying $P \supseteq \mathbb{B}$ and*

$$d^{\mathscr{H}}\left(T_\alpha^*, T_\alpha^P\right) \le \varepsilon. \tag{26}$$

In the remainder of the paper, we fix $\varepsilon > 0$ and choose a polytope P such that T_α^P and T_α^* satisfying (26). The polytope can be described as an interaction of a finite number of half spaces:

$$P := \left\{z \in \mathbb{R}^l : \left(a^i\right)^T, \quad z \le 1, i \in 1, \ldots, L\right\}, \tag{27}$$

where a^i for $i \in 1, \ldots, L$ are some unit vectors (i.e., $\|a^i\|_2 = 1$) and L is an integer number. For the convenience in the following analysis, we further add $2l$ half spaces

$$\begin{cases} \left(\beta^j\right)^T z \le 1, \\ -\left(\beta^j\right)^T z \le 1, \end{cases} \quad j = 1, \ldots, l, \tag{28}$$

to P, where β^j is the j-th column of the $l \times l$ identity matrix. This yields the following polytope:

$$\begin{aligned} P_0 &:= P \cap \left\{z \in \mathbb{R}^l : \left(\beta^j\right)^T z \le 1, -\left(\beta^j\right)^T z \le 1, \quad j = 1, \ldots, l\right\} \\ &= \left\{ z \in \mathbb{R}^l : \begin{array}{ll} \left(a^i\right)^T z \le 1, & i \in 1, \ldots, L; \\ \left(\beta^j\right)^T z \le 1, & j = 1, \ldots, l; \\ -\left(\beta^j\right)^T z \le 1, & j = 1, \ldots, l. \end{array} \right\} \end{aligned} \tag{29}$$

Denote by Ω the collection of the vectors a^i and $\pm\beta^j$ in P_0, i.e.,

$$\Omega := \left\{a^i : \quad i \in 1, \ldots, L\right\} \cup \left\{\pm\beta^j : \quad j \in 1, \ldots, l\right\}. \tag{30}$$

Clearly, P_0 still satisfies (26), i.e.,

$$d^{\mathscr{H}}\left(T_\alpha^*, T_\alpha^{P_0}\right) \le \varepsilon. \tag{31}$$

Let $N := |\Omega|$ and let M_{P_0} be the matrix with column vectors in Ω. Thus, P_0 can be written as

$$P_0 = \left\{z \in \mathbb{R}^l : \left(M_{P_0}\right)^T z \le e^N\right\}, \tag{32}$$

where e^N is the vector of one's in $\mathbb{R}^N$.

By replacing $\mathbb{B}$ by the above P_0, we obtain the following approximation of (3):

$$\min_x \left\{\|x\|_1 : b - Bx \in \varepsilon P_0, A^+ x \ge \alpha e_{J_+}, A^- x \le -\alpha e_{J_-}, A^0 x = 0\right\}, \tag{33}$$

and the solution set of (33) is

$$\begin{aligned} \left(T_\alpha^{P_0}\right)^* &= \left\{x : \|x\|_1 \le \left(\theta_\alpha^{P_0}\right)^*, b - Bx \in \varepsilon P_0, A^+ x \ge \alpha e_{J_+}, A^- x \le -\alpha e_{J_-}, A^0 x = 0\right\} \\ &= \left\{x : \|x\|_1 \le \left(\theta_\alpha^{P_0}\right)^*, r \in sP_0, r = b - Bx, (x,s) \in E_\alpha\right\}, \end{aligned} \tag{34}$$

where $(\theta_\alpha^{P_0})^*$ denotes the optimal value of (33). Since $\mathbb{B} \subseteq P_0$, then

$$\begin{gathered} \theta_\alpha^* \ge \left(\theta_\alpha^{P_0}\right)^*, \\ \left(T_\alpha^{P_0}\right)^* \subseteq T_\alpha^{P_0}, \\ T_\alpha^* \subseteq T_\alpha^{P_0}. \end{gathered} \tag{35}$$

3. Stability Analysis

The concept of range space property (RSP for short) was first introduced in [15] to develop a necessary and sufficient condition for uniform recovery of sparse signals via l_1-minimization. It was extended in [26] to weak RSP for developing stability theory of convex optimization algorithms. Recently, restricted RSP (RRSP) was

introduced to develop sign recovery condition for sparse signals through 1-bit measurement in [16, 25].

Definition 1 (weak RSP). Given a matrix $A \in \mathbb{R}^{m\times n}$, the transposed matrix A^T is said to possess the weak RSP order k, if for any two disjoint sets $S_1, S_2 \subseteq \{1, \ldots, n\}$ with $|S_1| + |S_2| \le k$, there exists a vector $\eta \in \mathscr{R}(A^T)$ such that

$$\begin{aligned} \eta_i &= 1, \quad \text{for } i \in S_1, \\ \eta_i &= -1, \quad \text{for } i \in S_2, \\ |\eta_i| &\le 1 \quad \text{for } i \notin S_1 \cup S_2. \end{aligned} \tag{36}$$

To investigate the stability of 1-bit compressed sensing involved noise constraints, the notion of weak RSP is needed to be extended to the following restricted weak RSP with respect to y.

Definition 2 (restricted weak RSP with respect to y). Given matrices $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{l\times n}$, and $y \in \{-1, 1, 0\}^m$, the pair (A^T, B^T) is said to satisfy the restricted weak RSP of order k with respect to y, if for any disjoint subsets S_1, S_2 of $\{1, \ldots, n\}$ with $|S_1| + |S_2| \le k$, there exists $\eta \in \mathscr{R}(A^T, B^T)$ such that

$$\eta = \left(A^T, B^T\right)\begin{pmatrix} w \\ h \end{pmatrix}, \tag{37}$$

where $w = (w^{(1)}, w^{(2)}, w^{(3)})^T \in \mathbb{R}_+^{|J_+|} \times \mathbb{R}_-^{|J_-|} \times \mathbb{R}^{|J_0|}$, $h \in \mathbb{R}^l$, and

$$\begin{aligned} \eta_i &= 1, \quad \text{for } i \in S_1, \\ \eta_i &= -1, \quad \text{for } i \in S_2, \\ |\eta_i| &\le 1, \quad \text{for } i \notin S_1 \cup S_2. \end{aligned} \tag{38}$$

Theorem 1. *Let $A \in \mathfrak{R}^{m\times n}$ and $B \in \mathfrak{R}^{l\times n}$ be given matrices and $b \in \mathfrak{R}^l$. Suppose that, for any given vector $y \in \{sign(Ax) \mid \|x\|_0 \le k\}$, the following holds: for any $x \in \mathfrak{R}^n$ satisfying $y = sign\,(Ax)$, there is a solution x^* of*

$$\begin{aligned} &\min_x \|x\|_1, \\ &\text{s.t. } A^+x \ge \alpha e_{J_+}, A^-x \le -\alpha e_{J_-}, A^0x = 0, \\ &Bx = b, \end{aligned} \tag{39}$$

where $\alpha > 0$ and A^+, A^0, and A^- are submatrices of A in which their rows are corresponding to index sets $J_+(y)$, $J_-(y)$, and $J_0(y)$, such that

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \le C\sigma_k(x)_1. \tag{40}$$

Here, C is a constant dependent only on the problem data (A, B, y, b). Then, (A^T, B^T) must satisfy the restricted weak RSP of order k with respect to y.

Proof. Let (S_1, S_2) be any pair of disjoint subsets of $\{1, \ldots, n\}$ with $|S_1| + |S_2| \le k$. To prove that (A^T, B^T) satisfies the restricted weak RSP of order k with respect to y, it is sufficient to show that there exists a vector $\eta \in \mathscr{R}(A^T, B^T)$ such that

$$\eta = \left(A^T, B^T\right)\begin{pmatrix} w \\ h \end{pmatrix}, \tag{41}$$

where $w := (w^{(1)}, w^{(2)}, w^{(3)})^T \in \mathbb{R}_+^{|J_+|} \times \mathbb{R}_-^{|J_-|} \times \mathbb{R}^{|J_0|}$, $h \in \mathbb{R}^l$, and

$$\begin{aligned} \eta_i &= 1, \quad \text{for } i \in S_1; \\ \eta_i &= -1, \quad \text{for } i \in S_2; \\ |\eta_i| &\le 1, \quad \text{for } i \notin S_1 \cup S_2. \end{aligned} \tag{42}$$

Take a k-sparse vector $\hat{x}$ in $\mathbb{R}^n$. Define

$$\begin{aligned} S_1 &:= \{i\colon \hat{x}_i > 0\}, \\ S_2 &:= \{i\colon \hat{x}_i < 0\}. \end{aligned} \tag{43}$$

Let $y := \text{sign}\,(A\hat{x})$. By assumption, there is a solution x^* of (39) such that

$$\left\| \frac{\hat{x}}{\|\hat{x}\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \le C\sigma_k(\hat{x})_1. \tag{44}$$

Since $\hat{x}$ is k-sparse, then $\sigma_k(\hat{x})_1 = 0$, which in turn implies $\hat{x}/\|\hat{x}\| = x^*/\|x^*\|$. So, sign $(\hat{x})$ = sign (x^*). This, together with (43), implies that

$$\begin{aligned} \{i\colon x_i^* > 0\} &= S_1, \\ \{i\colon x_i^* < 0\} &= S_2, \\ \{i\colon x_i^* = 0\} &= (S_1 \cup S_2)^c. \end{aligned} \tag{45}$$

Since x^* is a solution of linear programming (39), then KKT conditions hold; i.e., there exist $w = (w^{(1)}, w^{(2)}, w^{(3)})^T \in \mathbb{R}^{|J_+|} \times \mathbb{R}^{|J_-|} \times \mathbb{R}^{|J_0|}$ and $h \in \mathbb{R}^l$ such that

$$\eta := \begin{bmatrix} A^+ \\ A^- \\ A^0 \end{bmatrix}^T w + B^T h \in \partial\left\|x^*\right\|_1, \tag{46}$$

where $\partial\|x^*\|_1$ is the subgradient of the l_1-norm at x^*, i.e.,

$$\partial\|x^*\|_1 = \left\{ v \in \mathbb{R}^n : \begin{array}{rl} v_i = 1, & \text{for } x_i^* > 0; \\ v_i = -1, & \text{for } x_i^* < 0; \\ |v_i| \le 1, & \text{otherwise.} \end{array} \right\} \tag{47}$$

Hence, (46) ensures that

$$\begin{aligned} \eta_i &= 1, && \text{for } x_i^* > 0; \\ \eta_i &= -1, && \text{for } x_i^* < 0; \\ |\eta_i| &\le 1, && \text{for } x_i^* = 0. \end{aligned} \tag{48}$$

This together with (45) means that $\eta = A^T w + B^T h$ satisfies (42). Since S_1 and S_2 are arbitrary disjoint subsets of $\{1, \ldots, n\}$ with $|S_1| + |S_2| \le k$, we conclude that (A^T, B^T) satisfies the restricted weak RSP of order k with respect to y.

We now further show that the restricted weak RSP with respect to y is a sufficient condition for (3) to be stable. Firstly, for the approximation problem (33), let us introduce variables t, s to yield the following equivalent form:

$$\begin{aligned} \min_{(x,t,s)} \quad & e^T t, \\ \text{s.t.} \quad & x + t \ge 0, && -x + t \ge 0, \\ & -s \ge -\varepsilon, && M_{P_0}^T B x + s e^N \ge M_{P_0}^T b, \\ & A^+ x \ge \alpha e_{J_+}, && A^- x \le -\alpha e_{J_-}, \\ & A^0 x = 0, && (t, s) \ge 0. \end{aligned} \tag{49}$$

Recall that the solution set of (49) is given as (34). The above optimization problem is a linear programming problem, and the dual problem can be written as

$$\begin{aligned} \max_{w} \quad & -\varepsilon w_3 + b^T M_{P_0} w_4 + \alpha e_{J_+}^T w_5 - \alpha e_{J_-}^T w_6, \\ & w_1 - w_2 + B^T M_{P_0} w_4 + (A^+)^T w_5 + (A^-)^T w_6 + (A^0)^T w_7 = 0, \\ \text{s.t.} \quad & w_1 + w_2 \le e, -w_3 + (e^N)^T w_4 \le 0, \\ & w_1, w_2 \in \mathbb{R}_+^n, w_3 \in \mathbb{R}_+, w_4 \in \mathbb{R}_+^N, (w_5, w_6, w_7) \in \mathbb{R}_+^{|J_+|} \times \mathbb{R}_-^{|J_-|} \times \mathbb{R}^{|J_0|}. \end{aligned} \tag{50}$$

According to the dual theory on linear programming, the solution of (49) can be characterized by KKT conditions. □

Lemma 2. *x^* is a solution to the problem (33) if and only if $(x^*, t^*, s^*, w^*) \in \Theta$, where*

$$\Theta := \left\{ (x, t, s, w) \left| \begin{array}{c} -x - t \le 0, x - t \le 0, s \le \varepsilon, -M_{P_0}^T B x - s e^N \le -M_{P_0}^T b; \\ -A^+ x \le -\alpha e_{J_+}, A^- x \le -\alpha e_{J_-}, A^0 x = 0; \\ w_1 - w_2 + B^T M_{P_0} w_4 + (A^+)^T w_5 + (A^-)^T w_6 + (A^0)^T w_7 = 0; \\ w_1 + w_2 \le e, -w_3 + (e^N)^T w_4 \le 0; \\ e^T t = -\varepsilon w_3 + b^T M_{P_0} w_4 + \alpha e_{J_+}^T w_5 - \alpha e_{J_-}^T w_6; \\ (t, s) \ge 0, w_i \ge 0, \quad i = 1, \ldots, 5, w_6 \le 0. \end{array} \right. \right\} \tag{51}$$

For the convenience of notations, the set in (51) can be written equivalently as

$$\begin{aligned}
&\Theta = \{z = \{x, t, s, w\} \mid M'z \le p, M''z = q\}, \quad \text{where} \\
&p := \left(0, 0, \varepsilon, -M_{P_0}^T b, -\alpha e_{J_+}, -\alpha e_{J_-}, e, 0, 0, 0, 0, 0, 0, 0, 0, 0\right)^T, \\
&q := (0, 0, 0)^T,
\end{aligned} \tag{52}$$

$$\begin{aligned}
M' &:= \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \\ D_3 & 0 \\ 0 & D_4 \end{bmatrix}, \\
M'' &:= [M^*, M^{**}],
\end{aligned} \tag{53}$$

$$\begin{aligned}
D_1 &:= \begin{bmatrix} -I & -I & 0 \\ I & -I & 0 \\ 0 & 0 & 1 \\ -M_{P_0}^T B & 0 & -e^N \\ -A^+ & 0 & 0 \\ A^- & 0 & 0 \end{bmatrix}, \\
D_2 &:= \begin{bmatrix} I & I & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -I & \left(e^N\right)^T & 0 & 0 & 0 \end{bmatrix}, \\
D_3 &:= \begin{bmatrix} 0 & -I & 0 \\ 0 & 0 & -I \end{bmatrix}, \\
D_4 &:= \begin{bmatrix} -I_n & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -I_n & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -I_N & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -I_{|J_+|} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{|J_-|} & 0 \end{bmatrix}, \\
M^* &:= \begin{bmatrix} A^0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & e^T & 0 \end{bmatrix}, \\
M^{**} &:= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ I & -I & 0 & B^T M_{P_0} & (A^+)^T & (A^-)^T & \left(A^0\right)^T \\ 0 & 0 & \varepsilon & -b^T M_{P_0} & -\alpha\left(e_{J^+}\right)^T & \alpha\left(e_{J^-}\right)^T & 0 \end{bmatrix}.
\end{aligned} \tag{54}$$

The following two lemmas play a key role to establish the stability theory on 1-bit CS problem.

Lemma 3 (Hoffman's error bound). *Let $M' \in \mathbb{R}^{m\times q}$ and $M'' \in \mathbb{R}^{l\times q}$ be given matrices and*

$$\mathscr{F} := \{z \in \mathbb{R}^q \colon M'z \le p, M''z = q\}. \tag{55}$$

For any vector x in $\mathbb{R}^q$, there is a point $x^* \in \mathscr{F}$ such that

$$\left\|x - x^*\right\|_2 \le \sigma_{\infty,2}\left(M', M''\right)\left\|\begin{pmatrix} \left(M'x - p\right)^+ \\ M''x - q \end{pmatrix}\right\|_1, \tag{56}$$

where the constant $\sigma_{\infty,2}(M', M'')$ is referred to as Robinson's constant defined by M_1 and M_2.

Hoffman's error bound indicates that, for a linear system $\mathscr{F}$, the distance from a point in space to $\mathscr{F}$ can be measured in terms of Robinson's constant and quantity of the linear system being violated at this point.

Lemma 4 (see [25], Lemma 6.2.2). *Given three convex compact sets T_1, T_2, and T_3 satisfy $T_1 \subseteq T_2$ and $T_3 \subseteq T_2$, then*

$$\left\|x - \pi_{T_1}(x)\right\|_2 \le d^{\mathscr{H}}\left(T_1, T_2\right) + 2\|x - z\|_2, \quad \forall x \in \mathbb{R}^n, z \in T_3. \tag{57}$$

Inspired by [25, 26], we obtain the following result, which states that the restricted weak RSP with respect to y is a sufficient condition for the l_1-minimization (3) to be stable in sparse vector recovery.

Theorem 2. *Let the problem data $(A, B, \varepsilon, b, y)$ is given as (3) and rank $(A; B) = m + l$. Let $\varepsilon' > 0$ be any prescribed small number, and let P_0 be the polytope given in (29) satisfying (26). If $C^T = (A^T, B^T)$ satisfies the restricted weak RSP of order k with respect to y, then for any nonzero $x \in \mathbb{R}^n$, there is an optimal solution x^* of (3) such that*

$$\begin{aligned}
&\left\|\frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2}\right\|_2 \\
&\le \tau(x)\left(\varepsilon' + 2\gamma\{2\sigma_k(x)_1 + c\left(\|Bx - b\|_1 + \|Ax - \alpha y\|_1 + \varepsilon\right) + \left(\|b - Bx\|_2 - \varepsilon\right)^+\right. \\
&\left. + \left\|\left(-A^+x + \alpha e_{J_+}\right)^+\right\|_1 + \left\|\left(A^-x + \alpha e_{J_-}\right)^+\right\|_1 + \left\|A^0 x\right\|_1\}\right),
\end{aligned} \tag{58}$$

where $\alpha > 0$ is sufficient small, $c := \|(CC^T)^{-1}C\|_{\infty \longrightarrow 1}$, $\gamma := \sigma_{\infty,2}(M', M'')$ is the Robinson constant with (M', M'') given in (53), and

$$\tau(x) := \begin{cases} \dfrac{2}{\|x - x^*\|_2}, & \text{if } 0 \in [x, x^*], \\ \dfrac{1}{\operatorname{dist}(0, [x, x^*])}, & \text{if } 0 \notin [x, x^*]. \end{cases} \tag{59}$$

In particular, if x is a feasible solution of (3), then there is an optimal solution x^* of (1) such that

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \le \tau(x)\left(\varepsilon' + 2\gamma\{2\sigma_k(x)_1 + c(\|Bx - b\|_1 + \|Ax - \alpha y\|_1 + \varepsilon)\}\right). \tag{60}$$

Proof. Let $x \in \mathbb{R}^n$ be an arbitrary nonzero vector and P_0 be the fixed polytope given in (29) satisfying (26) in Lemma 1. The proof is divided into the following four steps. □

Step 1. (t, s, w). The first step is to construct t, s, w. Constructing (t, s). Let

$$\begin{aligned} t &:= |x|, \\ s &:= \left\| \left(M_{P_0}\right)^T (b - Bx) \right\|_\infty. \end{aligned} \tag{61}$$

The choice of (t, s) ensures

$$\begin{aligned} (-x - t)^+ &= 0, \\ (x - t)^+ &= 0, \\ \left(M_{P_0}^T (b - Bx) - e^N s\right)^+ &= 0. \end{aligned} \tag{62}$$

Let S be the support set of the k largest components of $|x|$. Define

$$\begin{aligned} S_1 &:= \{i\colon x_i > 0, \quad i \in S\}, \\ S_2 &:= \{i\colon x_i < 0, \quad i \in S\}. \end{aligned} \tag{63}$$

Clearly, $S_1 \cap S_2 = \varnothing$ and $S = S_1 \cup S_2$ with $|S_1 \cup S_2| = |S| \le k$. Let S_3 be the complementary set of S. Hence, S_1, S_2, and S_3 are disjoint. Since $C^T = (A^T, B^T)$ satisfies the restricted weak RSP of order k with respect to y, there exists a vector $\eta \in R(A^T, B^T)$ such that

$$\eta = A^T h^* + B^T v^* \tag{64}$$

for some $h^* = (h_1^*, h_2^*, h_3^*)^T \in \mathbb{R}_+^{|J_+|} \times \mathbb{R}_-^{|J_-|} \times \mathbb{R}^{|J_0|}$, $v^* \in \mathbb{R}^l$, and

$$\begin{aligned} \eta_i &= 1, \quad \text{for } i \in S_1, \\ \eta_i &= -1, \quad \text{for } i \in S_2, \\ |\eta_i| &\le 1, \quad \text{for } i \in S_3. \end{aligned} \tag{65}$$

Now, we construct a dual feasible solution $w = (w_1, \ldots, w_7)$.

Constructing (w_1, w_2, w_3). Set w_1, w_2, and w_3 as follows:

$$\begin{aligned} (w_1)_i &:= \begin{cases} 0, & i \in S_1, \\ 1, & i \in S_2, \\ \dfrac{(|\eta_i| - \eta_i)}{2}, & i \in S_3, \end{cases} \\ (w_2)_i &:= \begin{cases} 1, & i \in S_1, \\ 0, & i \in S_2, \\ \dfrac{(|\eta_i| + \eta_i)}{2}, & i \in S_3, \end{cases} \\ w_3 &:= \|v^*\|_1. \end{aligned} \tag{66}$$

Hence, (w_1, w_2) satisfies

$$\begin{aligned} w_1 + w_2 &\le e, \\ w_2 - w_1 &= \eta, \\ w_1, w_2 &\ge 0. \end{aligned} \tag{67}$$

Constructing w_4. We assume, without loss of generality, that the first l columns in M_{P_0} are β_j $(j = 1, \ldots, l)$ and the second l columns of M_{P_0} are $-\beta_j$ $(j = 1, \ldots, l)$. The component of w_4 is assigned as follows:

$$\begin{cases} (w_4)_j := v_j^*, & \text{if } v_j^* > 0, j = 1, \ldots, l; \\ (w_4)_{j+l} := -v_j^*, & \text{if } v_j^* < 0, j = 1, \ldots, l; \\ (w_4)_j := 0, & \text{otherwise.} \end{cases} \tag{68}$$

It follows from the choice of w_3 and w_4 that

$$\begin{aligned} M_{P_0} w_4 &= v^*, \\ \|w_4\|_1 &= \|v^*\|_1, \\ w_4 &\geq 0, \end{aligned} \tag{69}$$

$$\left(-w_3 + \left(e^N\right)^T w_4\right)^+ = \left(-\|v^*\|_1 + \left(e^N\right)^T w_4\right)^+ = \left(-\|v^*\|_1 + \|w_4\|_1\right) = 0. \tag{70}$$

Constructing (w_5, w_6, w_7). Let $(w_5, w_6, w_7) := h^*$. Clearly, $(w_5, w_6, w_7) \in \mathbb{R}_+^{|J_+|} \times \mathbb{R}_-^{|J_-|} \times \mathbb{R}^{|J_0|}$.

With the above choice of $w = (w_1, \ldots, w_7)$, it follows from (64)–(70) that

$$\begin{cases} w_1 - w_2 + B^T M_{P_0} w_4 + (A^+)^T w_5 + (A^-)^T w_6 + \left(A^0\right)^T w_7 = 0; \\ (w_1 + w_2 - e)^+ = 0, \left(-w_3 + \left(e^N\right)^T w_4\right)^+ = 0; \\ t^- = 0, s^- = 0, (w_i)^- = 0, \quad i = 1, \ldots, 5, (w_6)^+ = 0. \end{cases} \tag{71}$$

Step 2. Calculating $\|x - \overline{x}\|_2$, where $\overline{x}$ is a solution of (33) for $\alpha \in (0, \tilde{\alpha})$, and $\tilde{\alpha}$ satisfies $T_\beta^* \subseteq T^*$ for all $\beta \in (0, \tilde{\alpha})$ as required in Corollary 1.

Define

$$\begin{cases} \Lambda := e^T t + \varepsilon w_3 - b^T M_{P_0} w_4 - \alpha e_{J_+}^T w_5 + \alpha e_{J_-}^T w_6, \\ \Upsilon := (s - \varepsilon)^+. \end{cases} \tag{72}$$

For (x, t, s, w) where (t, s, w) is constructed as above, Lemma 3 ensures the existence of $(\overline{x}, t\overline{t}n, q\overline{s}h, \overline{w}) \in \Theta$ such that

$$\left\| \begin{pmatrix} x \\ t \\ s \\ w \end{pmatrix} - \begin{pmatrix} \overline{x} \\ \overline{t} \\ \overline{s} \\ \overline{w} \end{pmatrix} \right\|_2 \leq \gamma \left\| \begin{pmatrix} \Lambda \\ \Upsilon \\ \left(-A^+ x + \alpha e_{J_+}\right)^+ \\ \left(A^- x + \alpha e_{J_-}\right)^+ \\ A^0 x \\ (x - t)^+ \\ (-x - t)^+ \\ \left(\left(M_{P_0}\right)^T (b - Bx) - s e^N\right)^+ \\ (w_1 + w_2 - e)^+ \\ \left(-w_3 + \left(e^N\right)^T w_4\right)^+ \\ w_1 - w_2 + B^T M_{P_0} w_4 + (A^+)^T w_5 + (A^-)^T w_6 + \left(A^0\right)^T w_7 \\ \left(t^-, s^-, w_1^-, w_2^-, w_3^-, w_4^-, w_5^-, w_6^+\right) \end{pmatrix} \right\|_1, \tag{73}$$

where $\gamma := \sigma_{\infty,2}(M', M'')$ is Robinson's constant determined by (M', M'') given in (53). Since the vector (x, t, s, w) satisfies (62) and (71), the inequality (73) can be simplified to

$$\|(x, t, s, w) - (\overline{x}, \overline{t}, \overline{s}, \overline{w})\|_2 \leq \gamma \left\{ |\Lambda| + |\Upsilon| + \left\| \begin{pmatrix} \left(-A^+ x + \alpha e_{J_+}\right)^+ \\ \left(A^- x + \alpha e_{J_-}\right)^+ \\ A^0 x \end{pmatrix} \right\|_1 \right\}. \tag{74}$$

Since

$$\max_{1 \leq i \leq N} \left| \left[\left(M_{P_0}\right)^T (Bx - b) \right]_i \right| \leq \|Bx - b\|_2, \tag{75}$$

we have $s \leq \|b - Bx\|_2$ by (61). Therefore,

$$\Upsilon = (s - \varepsilon)^+ \leq \left(\|b - Bx\|_2 - \varepsilon\right)^+. \tag{76}$$

It follows from (69) that

$$\begin{aligned}\Lambda &= e^T t + \varepsilon w_3 - b^T v^* - \alpha e_{J_+}^T h_1^* + \alpha e_{J_-}^T h_2^*,\\ &= e^T t + \varepsilon w_3 - x^T B^T v^* + (Bx-b)^T v^* - x^T A^T h^* + (Ax-\alpha y)^T h^*,\\ &= e^T t + \varepsilon w_3 - x^T \eta + (Bx-b)^T v^* + (Ax-\alpha y)^T h^*,\end{aligned} \tag{77}$$

where the second step comes from the fact $y = (e_{J_+}, -e_{J_-}, 0)^T$ and the last step uses the fact $\eta = A^T h^* + B^T v^*$ by (64). Hence,

$$|\Lambda| \le \left|e^T t - x^T \eta\right| + \varepsilon\left|w_3\right| + \left|(Bx-b)^T v^*\right| + \left|(Ax-\alpha y)^T h^*\right|. \tag{78}$$

Firstly, we focus on each term of the right-hand side of the above inequality, respectively. Recall that $t = |x|$. Therefore,

$$\begin{aligned}\left|e^T t - x^T \eta\right| &= \left|e_S^T t_S + e_{S_3}^T t_{S_3} - x_S^T \eta_S - x_{S_3}^T \eta_{S_3}\right|\\ &= \left|e_{S_3}^T t_{S_3} - x_{S_3}^T \eta_{S_3}\right| \le \left|e_{S_3}^T t_{S_3}\right| + \left|x_{S_3}^T \eta_{S_3}\right|\\ &\le \left\|x_{S_3}\right\|_1 + \left\|x_{S_3}\right\|_1 \left\|\eta_{S_3}\right\|_\infty \le \left\|x_{S_3}\right\|_1 + \left\|x_{S_3}\right\|_1\\ &= 2\left\|x_{S_3}\right\|_1 = 2\sigma_k(x)_1,\end{aligned} \tag{79}$$

where the second equality is from (65). By using the restricted weak RSP of order k with respect to y, we have

$$\begin{aligned}\max\left\{\|v^*\|_1, \|h^*\|_1\right\} \le \left\|\begin{pmatrix} v^* \\ h^* \end{pmatrix}\right\|_1 &= \left\|\left(CC^T\right)^{-1} C\eta\right\|_1 \le \left\|\left(CC^T\right)^{-1} C\right\|_{\infty\longrightarrow 1} \|\eta\|_\infty\\ &\le \left\|\left(CC^T\right)^{-1} C\right\|_{\infty\longrightarrow 1} =: c.\end{aligned} \tag{80}$$

Hence,

$$\begin{cases} \left|(Bx-b)^T v^*\right| \le \|Bx-b\|_1 \left\|v^*\right\|_\infty \le \|Bx-b\|_1 \left\|v^*\right\|_1 \le c\|Bx-b\|_1,\\ \left|(Ax-\alpha y)^T h^*\right| \le \|Ax-\alpha y\|_1 \left\|h^*\right\|_\infty \le \|Ax-\alpha y\|_1 \left\|h^*\right\|_1 \le c\|Ax-\alpha y\|_1,\\ \varepsilon\left|w_3\right| = \varepsilon\left\|v^*\right\|_1 \le c\varepsilon. \end{cases} \tag{81}$$

It then follows from (78)–(81) that

$$|\Lambda| \le c\varepsilon + 2\sigma_k(x)_1 + c\left(\|Bx-b\|_1 + \|Ax-\alpha y\|_1\right), \tag{82}$$

together with (74) and (76) implies

$$\begin{aligned}&\|x-\overline{x}\|_2\\ &\le \gamma\{2\sigma_k(x)_1 + c\left(\|Bx-b\|_1 + \|Ax-\alpha y\|_1 + \varepsilon\right) + \left(\|b-Bx\|_2 - \varepsilon\right)^+\\ &+ \left\|\left(-A^+ x + \alpha e_{J_+}\right)^+\right\|_1 + \left\|\left(A^- x + \alpha e_{J_-}\right)^+\right\|_1 + \left\|A^0 x\right\|_1\}.\end{aligned} \tag{83}$$

Step 3. Calculating $\|x - x^*\|_2$, where x^* is a solution of (3). Recall three sets T_α^*, $T_\alpha^{P_0}$, and $(T_\alpha^{P_0})^*$, where T_α^* and $(T_\alpha^{P_0})^*$ are the solution of (18) and (33) (cf. (24) and (34)) and $T_\alpha^{P_0}$ is given as (25) with $P := P_0$. Clearly, $\overline{x} \in (T_\alpha^{P_0})^*$. Let x^* denote the projection of x onto T^*, i.e., $x^* = \pi_{T^*}(x)$. Since $T_\alpha^* \subseteq T_\alpha^{P_0}$ and $(T_\alpha^{P_0})^* \subseteq T_\alpha^{P_0}$ by (35), applying Lemma 4 with $T_1 := T_\alpha^*, T_2 := T_\alpha^{P_0}$ and $T_3 := (T_\alpha^{P_0})^*$, the definition of $\pi_T(x)$ and the fact $T_\alpha^* \subseteq T^*$ by Corollary 1 yields

$$\|x - x^*\|_2 = \|x - \pi_{T^*}(x)\|_2 \le \left\|x - \pi_{T_\alpha^*}(x)\right\|_2 \le d^{\mathscr{H}}\left(T_\alpha^*, T_\alpha^{P_0}\right) + 2\|x-\overline{x}\|_2, \tag{84}$$

which together with $d^{\mathscr{H}}(T_\alpha^*, T_\alpha^{P_0}) \le \varepsilon\prime$ by Lemma 1 implies

$$\left\|x - x^*\right\|_2 \le \varepsilon' + 2\|x-\overline{x}\|_2. \tag{85}$$

This combined with the inequality (83) gives

$$\begin{aligned}&\left\|x - x^*\right\|_2\\ &\le \varepsilon' + 2\gamma\{2\sigma_k(x)_1 + c\left(\|Bx-b\|_1 + \|Ax-\alpha y\|_1 + \varepsilon\right) + \left(\|b-Bx\|_2 - \varepsilon\right)^+\\ &+ \left\|\left(-A^+ x + \alpha e_{J_+}\right)^+\right\|_1 + \left\|\left(A^- x + \alpha e_{J_-}\right)^+\right\|_1 + \left\|A^0 x\right\|_1\}.\end{aligned} \tag{86}$$

Step 4. Calculating $\|(x/\|x\|_2) - (x^*/\|x^*\|_2)$. Note first that $x^* \neq 0$ due to $J_+ \cup J_- \neq \varnothing$. Consider the following two cases:

(i) If $0 \in [x, x^*]$, since $x, x^* \neq 0$, then $x = \alpha x^*$ for some $\alpha \neq 0$. Hence,

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 = \left\| \frac{ax^*}{\|ax^*\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \leq \frac{2}{\|x - x^*\|_2} \|x - x^*\|_2. \tag{87}$$

(ii) If $0 \notin [x, x^*]$, let $f(z) := z/\|z\|_2$ as $z \neq 0$. Then,

$$\nabla f(z) = \frac{I - (z/\|z\|_2)(z/\|z\|_2)^T}{\|z\|_2}, \tag{88}$$

which implies $\|\nabla f(z)\|_2 = 1/\|z\|_2$ since eigenvalues of $I - (z/\|z\|_2)(z/\|z\|_2)^T$ are 0 and 1 with multiplicity $n-1$. Thus,

$$\begin{aligned} f(x) - f(x^*) &= \int_0^1 \nabla f(x^* + t(x - x^*))(x - x^*) \mathrm{d}t \\ &\leq \int_0^1 \|\nabla f(x^* + t(x - x^*))\|_2 \|(x - x^*)\|_2 \mathrm{d}t \\ &\leq \frac{1}{\operatorname{dist}(0, [x, x^*])} \|x - x^*\|_2, \end{aligned} \tag{89}$$

where the last inequality is due to the fact for any $t \in [0, 1]$,

$$\|\nabla f(x^* + t(x - x^*))\|_2 = \frac{1}{\|x^* + t(x - x^*)\|_2} \leq \frac{1}{\operatorname{dist}(0, [x, x^*])}. \tag{90}$$

Combining (87) and (89) together yields

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \leq \tau(x) \|x - x^*\|_2, \tag{91}$$

where

$$\tau(x) := \begin{cases} \dfrac{2}{\|x - x^*\|_2}, & \text{if } 0 \in [x, x^*], \\ \dfrac{1}{\operatorname{dist}(0, [x, x^*])}, & \text{if } 0 \notin [x, x^*]. \end{cases} \tag{92}$$

This together with (86) results in (58).

If x is the feasible solution of (3), then $(\|b - Bx\|_2 - \varepsilon)^+ = 0$ and

$$\left\| \left(-A^+ x + \alpha e_{J_+}\right)^+ \right\|_1 = \left\| \left(A^- x + \alpha e_{J_-}\right)^+ \right\|_1 = \|A^0 x\|_1 = 0, \tag{93}$$

as $\alpha > 0$ is sufficiently small, which further implies

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \leq \tau(x) \left(\varepsilon' + 2\gamma\{2\sigma_k(x)_1 + c(\|Bx - b\|_1 + \|Ax - \alpha y\|_1 + \varepsilon)\}\right). \tag{94}$$

We now further show that the restricted weak RSP with respect to y is also a sufficient condition for the l_1-minimization problem if the noise does not exist, i.e., $\varepsilon = 0$. It should be noticed that, in this case, the constraint $Bx = b$ is linear, and hence, it is unnecessary to further introduce a polytope. Thus, the problem (3) and its relaxed problem (49) reduces to

$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{s.t.} \quad & \operatorname{sign}(Ax) = y, \\ & Bx = b, \end{aligned} \tag{95}$$

$$\begin{aligned} \min_{x,t} \quad & e^T t, \\ & x + t \geq 0, -x + t \geq 0, \\ \text{s.t.} \quad & A^+ x \geq \alpha e_{J_+}, A^- x \leq -\alpha e_{J_-}, \\ & A^0 x = 0, Bx = b, t \geq 0. \end{aligned} \tag{96}$$

The dual problem is given as

$$\begin{aligned} \max_{w} \quad & \alpha e_{J_+}^T w_3 - \alpha e_{J_-}^T w_4 + b^T w_6, \\ & w_1 - w_2 + (A^+)^T w_3 + (A^-)^T w_4 + \left(A^0\right)^T w_5 + B^T w_6 = 0, \\ \text{s.t.} \quad & w_1 + w_2 \leq e, (w_1, w_2) \geq 0, w_6 \in \mathbb{R}^l, \\ & (w_3, w_4, w_5) \in \mathbb{R}_+^{|J_+|} \times \mathbb{R}_-^{|J_-|} \times \mathbb{R}^{|J_0|}. \end{aligned} \tag{97}$$

Similarly, according to the dual theory of linear programming, x^* is a solution to the problem (96) if and only if there exists $(x^*, t^*, s^*, w^*) \in \overline{\Theta}$, where

$$\overline{\Theta} := \left\{ (x, t, w) \left| \begin{array}{c} x \leq t, -x \leq t, -A^+ x \leq -\alpha e_{J_+}, A^- x \leq -\alpha e_{J_-}, A^0 x = 0, Bx = b; \\ w_1 - w_2 + \left(A^+\right)^T w_3 + (A^-)^T w_4 + \left(A^0\right)^T w_5 + B^T w_6 = 0, w_1 + w_2 \leq e; \\ e^T t = \alpha e_{J_+}^T w_3 - \alpha e_{J_-}^T w_4 + b^T w_6, (w_1, w_2, t) \geq 0, w_3 \geq 0, w_4 \leq 0. \end{array} \right. \right. \tag{98}$$

The set $\overline{\Theta}$ can be written equivalently as

$$\overline{\Theta} = \{z = (x, t, w) \mid \overline{M}' z \le \overline{p}, \overline{M}'' z = \overline{q}\}, \tag{99}$$

where $\overline{p} := (0, 0, -\alpha e_{J_+}, -\alpha e_{J_-}, e, 0, 0, 0, 0, 0,)$, $\overline{q} := (0, b, 0, 0)$,

$$\begin{aligned}
\overline{M}' &:= \begin{bmatrix} \overline{D}_1 & 0 \\ 0 & \overline{D}_2 \\ \overline{D}_3 & 0 \\ 0 & \overline{D}_4 \end{bmatrix}, \\
\overline{M}'' &:= \left[\overline{M}^*, \overline{M}^{**}\right],
\end{aligned} \tag{100}$$

$$\begin{aligned}
\overline{D}_1 &:= \begin{bmatrix} I & -I \\ -I & -I \\ -A^+ & 0 \\ A^- & 0 \end{bmatrix}, \\
\overline{D}_2 &:= \begin{bmatrix} I & I & 0 & 0 & 0 & 0 \end{bmatrix}, \\
\overline{D}_3 &:= \begin{bmatrix} 0 & -I \end{bmatrix}, \\
\overline{D}_4 &:= \begin{bmatrix} -I_n & 0 & 0 & 0 & 0 & 0 \\ 0 & -I_n & 0 & 0 & 0 & 0 \\ 0 & 0 & -I_{|J_+|} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{|J_-|} & 0 & 0 \end{bmatrix}, \\
\overline{M}^* &:= \begin{bmatrix} A^0 & 0 \\ B & 0 \\ 0 & 0 \\ 0 & e^T \end{bmatrix}, \\
\overline{M}^{**} &:= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ I & -I & (A^+)^T & (A^-)^T & (A^0)^T & B^T \\ 0 & 0 & -\alpha e_{J_+} & \alpha e_{J_-} & 0 & -b^T \end{bmatrix}.
\end{aligned} \tag{101}$$

Following the similar argument as given in Theorem 2, we can obtain the following result.

Theorem 3. *Let the problem data (A, B, b, y) is given as (95) and the matrix $C = (A^T, B^T)^T \in \mathbb{R}^{(m+l)\times n}$ with full row rank. If $C^T = (A^T, B^T)$ satisfies the restricted weak RSP of order k with respect to y, then for any $x \in \mathbb{R}^n$, there is an optimal solution x^* of (3) such that*

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \le 2\gamma\tau(x)\left\{2\sigma_k(x)_1 + c\|Ax - \alpha y\|_1 + \left\|\left(-A^+x + \alpha e_{J_+}\right)^+\right\|_1 + \left\|\left(A^-x + \alpha e_{J_-}\right)^+\right\|_1 + \left\|A^0 x\right\|_1\right\}, \tag{102}$$

where $\alpha > 0$ is sufficiently small, $c := \|(CC^T)^{-1}C\|_{\infty \longrightarrow 1}$, and $\gamma := \sigma_{\infty,2}(\overline{M}', \overline{M}'')$ is the Robinson constant with $(\overline{M}', \overline{M}'')$ given in (100). In particular, if x is a feasible solution of (3), then there is an optimal solution x^* of (3) such that

$$\left\| \frac{x}{\|x\|_2} - \frac{x^*}{\|x^*\|_2} \right\|_2 \le 2\gamma\tau(x)\{2\sigma_k(x)_1 + c\|Ax - \alpha y\|_1\}. \tag{103}$$

The following result shows that the property of restricted weak RSP with respect to y is the mildest condition to ensure the stability of l_1-minimization problem with any given measurement vector $y = (e_{J_+}, -e_{J_-}, 0) \in \{\text{sign}(Ax) : \|x\|_0 \le k\}$.

Corollary 2. *Let the problem data (A, B, b, y) be given as (95) and $C = (A^T, B^T)^T \in \mathbb{R}^{(m+l)\times n}$ be a matrix with full row rank. Then, the 1-bit CS problem*

$$\begin{aligned} \min \quad & \|x\|_1 \\ \text{s.t.} \quad & A^+x \geq \alpha e_{J_+}, A^-x \leq -\alpha e_{J_-}, A^0x = 0, \\ & Bx = b, \end{aligned} \tag{104}$$

is stable for all $y \in \{sign\ (Ax)\colon \|x\|_0 \leq k\}$ *if and only if* C^T *satisfies restricted weak RSP of order* k *with respect to* y.

Proof. Following the argument given in Theorem 2, we know that the restricted weak RSP of order k of C^T with respect to y is a sufficient condition for l_1-minimization problem (104) to be stable.

On the contrary, Theorem 1 claims that if the l_1-minimization problem is stable for any given $y \in \{\text{sign}\ (Ax)\colon \|x\|_0 \leq k\}$, then the matrix C^T must satisfy the restricted weak RSP of the order k with respect to y. □

4. Conclusions

In this paper, the stability theory for 1-bit CS with quadratic constraint is established. In the analysis, it is essential to use the duality theory of linear programming, Hoffman error bound, and the fact that the ball constraint via Euclidean norm can be approximated by polytopes to any level of accuracy. An interesting and challenging topic is to further study the stability theory for 1-bit CS with other norms, e.g., p-norm, particularly as $p \in (0, 1)$. In this case, the nonconvex structure of p-norm requires us to adopt the error bounded theory (also called metric subregularity) for nonlinear systems, instead of linear system used in this paper.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (11771255 and 11801325), Young Innovation Teams of Shandong Province (2019KJI013), Program of Science and Technology Activities for Overseas Students in Henan Province in 2020, and Nanhu Scholars Program for Young Scholars of Xinyang Normal University.

References

[1] N. Bi and J. Tan, "Characterization of ℓ_1 minimizer in one-bit compressed sensing," *Analysis and Applications*, vol. 17, no. 6, pp. 1005–1021, 2019.

[2] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[3] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[4] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[7] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[8] I. Daubechies, R. Devore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[9] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin, "Certifying the restricted isometry property is hard," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3448–3450, 2013.

[10] R. Baraniuk, M. Davenport, R. Devore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[11] J. C. Emmanuel, "The restricted isometry property and its implications for compressed sensing," *Comptes Rendus Mathématique*, vol. 346, pp. 589–592, 2008.

[12] J. C. Emmanuel, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, pp. 59–73, 2011.

[13] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[14] C. Yang, X. Shen, H. Ma, Y. Gu, and H. C. So, "Sparse recovery conditions and performance bounds for ℓ_p-minimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 19, pp. 5014–5028, 2018.

[15] Y.-B. Zhao, "RSP-based analysis for sparsest and least ℓ_1-norm solutions to underdetermined linear systems," *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5777–5788, 2013.

[16] Y. Zhao and C. Xu, "1-Bit compressive sensing: reformulation and RRSP-based sign recovery theory," *Science China Mathematics*, vol. 59, no. 10, pp. 2049–2074, 2016.

[17] J. Liu, J. Jin, and Y. Gu, "Robustness of sparse recovery via," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 3996–4014, 2015.

[18] C. Z. Yang, X. Y. Shen, H. B. Ma, B. D. Chen, Y. T. Gu, and H. C. So, "Weakly convex regularized robust sparse recovery methods with theoretical guarantees," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 5046–5061, 2019.

[19] R. Kueng and P. Jung, "Robust nonnegative sparse recovery and the nullspace property of 0/1 measurements," *IEEE Transactions on Information Theory*, vol. 64, pp. 689–703, 2018.

[20] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Proceedings of the 2008 42nd Annual Conference on Information Sciences and Systems*, IEEE, Princeton, NJ, USA, March 2008.

[21] U. Ayaz, S. Dirksen, and H. Rauhut, "Uniform recovery of fusion frame structured sparse signals," *Applied and Computational Harmonic Analysis*, vol. 41, pp. 341–361, 2014.

[22] R. Saab, R. Wang, and O. Yilmaz, "Quantization of compressive samples with stable and robust recovery," *Applied and Computational Harmonic Analysis*, vol. 44, pp. 123–143, 2018.

[23] Q. Sun, "Recovery of sparsest signals via l_q-minimization," *Applied and Computational Harmonic Analysis*, vol. 32, pp. 329–341, 2012.

[24] J. L. Xu and Y. B. Zhao, "Stability analysis for a class of sparse optimization problems," *Optimization Methods and Software*, vol. 35, pp. 836–854, 2020.

[25] Y. B. Zhao, *Sparse Optimization Theory and Methods*, CRC Press, Taylor, Francis Group, Boca Raton, FL, USA, 2018.

[26] Y. B. Zhao, H. Y. Jiang, and Z. Q. Luo, "Weak stability of l_1-minimization methods in sparse data reconstruction," *Mathematics of Operations Research*, vol. 44, pp. 173–195, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 7487383, 14 pages
https://doi.org/10.1155/2020/7487383

Research Article

Shrinking Projection Methods for Accelerating Relaxed Inertial Tseng-Type Algorithm with Applications

Hasanen A. Hammad,[1] **Habib ur Rehman**,[2] **and Manuel De la Sen**[3]

[1]*Department of Mathematics, Faculty of Science, Sohag University, Sohag 82524, Egypt*
[2]*Department of Mathematics, Monglkuts University of Technology, Bangkok 10140, Thailand*
[3]*Institute of Research and Development of Processes University of the Basque Country, Leioa (Bizkaia) 48940, Spain*

Correspondence should be addressed to Hasanen A. Hammad; h.elmagd89@gmail.com

Received 8 August 2020; Revised 17 September 2020; Accepted 7 October 2020; Published 6 November 2020

Academic Editor: Guoqiang Wang

Our main goal in this manuscript is to accelerate the relaxed inertial Tseng-type (RITT) algorithm by adding a shrinking projection (SP) term to the algorithm. Hence, strong convergence results were obtained in a real Hilbert space (RHS). A novel structure was used to solve an inclusion and a minimization problem under proper hypotheses. Finally, numerical experiments to elucidate the applications, performance, quickness, and effectiveness of our procedure are discussed.

1. Introduction

The standard form of the variational inclusion problem (VIP) on a RHS $\daleth$ is

$$0 \in (¥ + \Upsilon)\vartheta^*, \tag{1}$$

where ϑ^* is the unknown point that we need to find, for an operator $¥: \daleth \longrightarrow \daleth$ and a set-valued operator $\Upsilon: \daleth \longrightarrow 2^{\daleth}$. VIP is a frequent problem in the optimization field, which has a lot of applications in many areas, including equilibrium, machine learning, economics, engineering, image processing, and transportation problems [1–16].

The vintage technique to solve problem (1) which is denoted by $(¥ + \Upsilon)^{-1}(0)$ is the forward-backward splitting method [17–22] which is defined as follows: $\vartheta_1 \in \daleth$ and

$$\vartheta_{n+1} = (I + \ell\Upsilon)^{-1}(I - \ell¥)\vartheta_n, \quad n \geq 1, \tag{2}$$

where $\ell > 0$. In (2), each step of iterates includes only the forward step ¥ and the backward step Υ, but not $¥ + \Upsilon$. This technique involves the proximal point algorithm [23–25] and the gradient method [26–28] as special cases.

In a RHS, nice splitting iterative procedures presented by Lions and Mercier [29] are shown as follows:

$$\vartheta_{n+1} = \left(2J_\ell^{¥} - I\right)\left(2J_\ell^{\Upsilon} - I\right)\vartheta_n, \quad n \geq 1, \tag{3}$$

and

$$\vartheta_{n+1} = J_\tau^{¥}\left(2J_\ell^{\Upsilon} - I\right)\vartheta_n + \left(I - J_\ell^{\Upsilon}\right)\vartheta_n, \quad n \geq 1, \tag{4}$$

where $J_\ell^{\mathfrak{R}} = (I + \ell\mathfrak{R})^{-1}$. Permanently, two algorithms are weakly convergent [30], knowing that algorithm (3) is called Peaceman–Rachford algorithm [19] and scheme (4) is called Douglas–Rachford algorithm [31].

A lot of works are concerned with problem (1) for accretive operators and two monotone operators, for instance, a stationary solution to the initial-valued problem of the evolution equation

$$0 \in \frac{\partial \varpi}{\partial t} - \Xi\varpi, \varpi(0) = \varpi_\circ \tag{5}$$

can be adjusted as (1) when the governing maximal monotone $\Xi = ¥ + \Upsilon$ [29].

[1] is used to solve a minimization problem as follows:

$$\min_{\vartheta \in \daleth} \beth(\vartheta) + \sigma(\vartheta), \tag{6}$$

where $\beth, \sigma: \daleth \longrightarrow (-\infty, \infty]$ are proper and lower semi-continuous convex functions such that $\beth$ is differentiable with L-Lipschitz gradient, and the proximal mapping of σ is

$$\vartheta \longmapsto \arg\min_{\omega\in\daleth} \sigma(\omega) + \frac{\|\vartheta-\omega\|^2}{2\ell}. \tag{7}$$

In particular, if $¥ = \nabla\beth$ and $\Upsilon = \partial\sigma$, where $\nabla\beth$ is the gradient of $\beth$ and $\partial\sigma$ is the subdifferential of σ which takes the form $\partial\sigma(\vartheta) = \{\lambda \in \daleth: \sigma(\omega) \geq \sigma(\vartheta) + \langle\lambda, \omega-\vartheta\rangle \forall\omega \in \daleth\}$, problem (1) becomes (6), and (3) becomes

$$\vartheta_{n+1} = \operatorname{prox}_{\ell\sigma}(\vartheta_n - \ell\nabla\beth(\vartheta_n)), \quad n \geq 1, \tag{8}$$

where $\ell > 0$ is the stepsize and $\operatorname{prox}_{\ell\sigma} = (I + \ell\,\partial\sigma)^{-1}$ is the proximity operator of σ.

The concept of merging the inertial term with the backward step was initiated by Alvarez and Attouch [32] and studied extensively in [33, 34]. For maximal monotone operators, it was called the inertial proximal point (IPP) algorithm, and they defined it by

$$\begin{cases} \mathfrak{I}_n = \vartheta_n + \Lambda_n(\vartheta_n - \vartheta_{n-1}), \\ \vartheta_{n+1} = (I + \ell_n\Upsilon)^{-1}\mathfrak{I}_n, \; n \geq 1. \end{cases} \tag{9}$$

It was proved that if $\{\ell_n\}$ is nondecreasing and $\{\Lambda_n\} \subset [0,1)$ with

$$\sum_{n=1}^{\infty} \Lambda_n \|\vartheta_n - \vartheta_{n-1}\|^2 < \infty, \tag{10}$$

then algorithm (9) converges weakly to zero of Υ. In particular, condition (10) is true for $\Lambda_n < 1/3$. Here, Λ_n is an extrapolation factor, and the inertia is represented by the term $\Lambda_n(\vartheta_n - \vartheta_{n-1})$. Note that the inertial term improves the performance of the procedure and has good convergence results [35–37].

Inertial term was merged with forward-backward algorithm by authors [38]. They added Lipschitz-continuous, a single-valued, cocoercive operator ¥ into the IPP algorithm:

$$\begin{cases} \mathfrak{I}_n = \vartheta_n + \Lambda_n(\vartheta_n - \vartheta_{n-1}), \\ \vartheta_{n+1} = (I + \ell_n\Upsilon)_n^{-1}(\mathfrak{I}_n - \ell_n ¥\mathfrak{I}_n), \quad n \geq 1. \end{cases} \tag{11}$$

Via assumption (10), provided $\ell_n < 2/L$ with L, the Lipschitz constant of ¥, they obtained a weak convergence result. Note that, for $\Lambda_n > 0$, algorithm (11) does not take the form of (2), in spite of ¥ is still evaluated at the points $¥_n$.

Relaxation techniques and inertial effects have many advantages in solving monotone inclusion and convex optimization problems; this effect appeared in several names such as relaxed inertial proximal method, relaxed inertial forward-backward method, and relaxed inertial Douglas–Rachford algorithm; for more details, refer to [22, 24, 39–44].

Abubakar et al. [45] introduced the RITT method as follows:

$$\begin{cases} \mathfrak{I}_n = \vartheta_n + \Lambda(\vartheta_n - \vartheta_{n-1}), \\ \psi_n = (1 + \ell_n\Upsilon)^{-1}(1 - \ell_n ¥)\mathfrak{I}_n, \\ \phi_{n+1} = (1-\beta)\mathfrak{I}_n + \beta\psi_n + \beta\ell_n(¥\mathfrak{I}_n - ¥\psi_n), \; n \geq 1, \end{cases} \tag{12}$$

where Λ and β are extrapolation and relaxation parameters, respectively. Under this algorithm, they discussed the weak convergence to the solution point of VIP (1) and the problem of image recovery. Note that the extrapolation step works to accelerate but not for the desired acceleration.

The concept of the SP method was discussed by Takahashi et al. [46] as in the following algorithm:

$$\begin{cases} \vartheta_0 \in \daleth \text{ be arbitrarily fixed}, \\ C_1 = C, \vartheta_1 = P_{C_1}\vartheta_0, \\ \omega_n = \Lambda_n\vartheta_n + (1-\Lambda_n)\hbar_n\vartheta_n, \\ C_n = \{\eta \in C: \|\omega_n - \eta\| \leq \|\vartheta_n - \eta\|\}, \\ \vartheta_{n+1} = P_{C_{n+1}}\vartheta_0. \end{cases} \tag{13}$$

They just selected one closed convex (CC) set for a family of nonexpansive mappings $\{\hbar_n\}$ to modify Mann's iteration method [47] and proved that the sequence $\{\vartheta_n\}$ converges strongly to $P_{\operatorname{Fix}(\hbar)}\vartheta_0$, provided $\Lambda_n \leq e$ for all $n \geq 1$ and for some $0 < e < 1$.

In 2019, Yang and Liu [48] selected the stepsize sequence for the iterative algorithm for monotone variational inequalities, which are based on Tseng's extragradient method and Moudafi viscosity scheme that does not require either the knowledge of the Lipchitz constant of the operator or additional projections.

With the incorporation of results of [45, 46, 48], we accelerate RITT algorithm by adding the SP method to algorithm (12). In a RHS, strong convergence results are given under a proposed algorithm. As applications, our algorithm was used to find the solution to a VIP and minimization problem under certain conditions. Eventually, numerical experiments to illustrate the applications, performance, acceleration, and effectiveness of the proposed algorithm are presented.

2. Preparatory Lemmas and Definitions

Suppose that C is a nonempty closed convex subset (CCS) of a RHS $\daleth$; we shall refer to " $\longrightarrow$ " as the strong convergence, and P_C: $\daleth \longrightarrow C$ is the nearest point projection, that is, for all $\vartheta \in \daleth$ and $\omega \in C$, $\|\vartheta - P_C\vartheta\| \leq \|\vartheta - \omega\|$. P_C is called the metric projection. It is obvious that P_C verifies the following inequality:

$$\|P_C\vartheta - P_C\omega\|^2 \leq \langle P_C\vartheta - P_C\omega, \vartheta - \omega\rangle, \tag{14}$$

for all $\vartheta, \omega \in \daleth$. In other words, the metric projection P_C is firmly nonexpansive. Hence, $\langle \vartheta - P_C\vartheta, \omega - P_C\omega\rangle \leq 0$ holds for all $\vartheta \in \daleth$ and $\omega \in C$, see [49, 50].

The following inequality holds in a HS [51]:

$$\|l \pm m\|^2 = \|l\|^2 + \|m\|^2 \pm 2\langle l, m\rangle, \tag{15}$$

for all $l, m \in \daleth$.

Lemma 1 (see [52]). *Let C be a nonempty CCS of a RHS $\daleth$. For each $\vartheta, \omega, \upsilon \in \daleth$ and $\in \mathbb{R}$, the following set is closed and convex:*

$$\left\{\eta \in C\colon \|\omega - \eta\|^2 \le \|\vartheta - \eta\|^2 + \langle \upsilon, \eta\rangle + \delta\right\}. \tag{16}$$

Lemma 2 (see [38]). *Let C be a nonempty CCS of a RHS $\daleth$ and $P_C\colon \daleth \longrightarrow C$ be the metric projection. Then,*

$$\left\|\omega - P_C\vartheta\right\|^2 + \left\|\vartheta - P_C\vartheta\right\|^2 \le \|\vartheta - \omega\|^2, \tag{17}$$

for all $\vartheta \in \daleth$ and $\omega \in C$.

Definition 1. Suppose that $D(¥) \subset \daleth$ and $R(¥) \subset \daleth$ are the domain and the range of an operator ¥, respectively. For all $\vartheta, \omega \in D(¥)$, an operator ¥ is called

(1) Monotone if

$$\langle \vartheta - \omega, ¥\vartheta - ¥\omega\rangle \ge 0. \tag{18}$$

(2) L–Lipschitz if

$$\|¥\vartheta - ¥\omega\| \le L\|\vartheta - \omega\|. \tag{19}$$

(3) β–Strongly monotone if there exists $\beta > 0$ such that

$$\langle \vartheta - \omega, ¥\vartheta - ¥\omega\rangle \ge \beta\|\vartheta - \omega\|^2. \tag{20}$$

(4) Λ–Inverse strongly monotone (Λ–ism) if there exists $\Lambda > 0$ such that

$$\langle \vartheta - \omega, ¥\vartheta - ¥\omega\rangle \ge \Lambda\|¥\vartheta - ¥\omega\|^2. \tag{21}$$

Lemma 3 (see [44]). *Let $\daleth$ be a RHS, ¥: $\gimel \longrightarrow \gimel$ be an Λ–ism operator, and $\Upsilon\colon \daleth \longrightarrow 2^{\daleth}$ be a maximal monotone operator. For each $\ell > 0$, we define*

$$\mho_\ell = J_\ell^{\Upsilon}(I - \ell¥) = (I + \ell\Upsilon)^{-1}(I - \ell¥). \tag{22}$$

Then, we get

(i) For $\ell > 0$, $fix(\mho_\ell) = (¥ + \Upsilon)^{-1}(0)$

(ii) For $0 < s \le \ell$ and $\vartheta \in \daleth$, $\|\vartheta - \mho_s\vartheta\| \le 2\|\vartheta - \mho_\ell\vartheta\|$

Lemma 4. *Let $\daleth$ be a RHS, ¥: $\daleth \longrightarrow \daleth$ be an Λ–ism operator, and $\Upsilon\colon \daleth \longrightarrow 2^{\daleth}$ be a maximal monotone operator. For each $\ell > 0$, we have*

$$\left\|\mho_\ell\vartheta - \mho_\ell\omega\right\|^2 \le \|\vartheta - \omega\|^2 - \ell(2\Lambda - \ell)\|¥\vartheta - ¥\omega\|^2, \tag{23}$$

for all $\vartheta, \omega \in \daleth$.

Proof. For all $\vartheta, \omega \in \daleth$, we get

$$\begin{aligned}
\left\|\mho_\ell\vartheta - \mho_\ell\omega\right\|^2 &= \left\|J_r^{\Upsilon}(I - \ell¥)\vartheta - J_r^{\Upsilon}(I - \ell¥)\omega\right\|^2 \\
&\le \|(I - \ell¥)\vartheta - (I - \ell¥)\omega\|^2 \\
&= \|(\vartheta - \omega) - \ell(¥\vartheta - ¥\omega)\|^2 \\
&= \|\vartheta - \omega\|^2 - 2\ell\langle \vartheta - \omega, ¥\vartheta - ¥\omega\rangle + \ell^2\|¥\vartheta - ¥\omega\|^2 \\
&\le \|\vartheta - \omega\|^2 - 2\ell\Lambda\|¥\vartheta - ¥\omega\|^2 + \ell^2\|¥\vartheta - ¥\omega\|^2 \\
&= \|\vartheta - \omega\|^2 - \ell(2\Lambda - \ell)\|¥\vartheta - ¥\omega\|^2.
\end{aligned} \tag{24}$$

The proof is ended. □

3. Shrinking Projection Relaxed Inertial Tseng-Type Algorithm

We provide a method consisting of the forward-backward splitting method with an inertial factor and an explicit stepsize formula, which are being used to ameliorate the convergence average of the iterative scheme and to make the manner independent of the Lipschitz constants. The detailed method is provided in Algorithm 1.

Note that

(i) Since ¥ is an Λ–ism operator, it is a Lipschitz function with a constant L, $¥\mathfrak{I}_n \ne ¥\psi_n$, and we get

$$\frac{\rho\left\|\mathfrak{I}_n - \psi_n\right\|}{\left\|¥\mathfrak{I}_n - ¥\psi_n\right\|} \ge \frac{\rho}{L}. \tag{25}$$

It is obvious for $¥\mathfrak{I}_n = ¥\psi_n$ that inequality (25) is satisfied. Hence, it follows that $\ell_n \ge \min\{(\rho/L), \ell_0\}$. This implies that the generated sequence $\{\ell_n\}$ is bounded below by $\min\{(\rho/L), \ell_0\}$, i.e., $\{\ell_n\}$ is monotonically decreasing.

(ii) By (i) and (25), we have

$$\ell_{n+1}\left\|¥\mathfrak{I}_n - ¥\psi_n\right\| \le \rho\left\|\mathfrak{I}_n - \psi_n\right\|, \tag{26}$$

i.e., the update (28) is well defined.

(iii) If we delete the shrinking projection term from our algorithm, we get the algorithms of the papers [22, 45, 53].

Theorem 1. *Let $\daleth$ be a RHS and the operators ¥: $\daleth \longrightarrow \daleth$ be Λ–ism on $\daleth$, and $\Upsilon\colon \daleth \longrightarrow 2^{\daleth}$ is maximally monotone. If feasible set $\Omega = (¥ + \Upsilon)^{-1}(0)$ of (1) is a nonempty CCS of a RHS $\daleth$, then the sequence $\{\vartheta_n\}$ generated by Algorithm 1 converges strongly to a point $\tau = P_\Omega(\vartheta_1)$, provided that*

(i) $0 < \liminf_{n\longrightarrow\infty}\ell\ell_n \le \limsup_{n\longrightarrow\infty}\ell_n < 2\Lambda$.

(ii) $\lim_{n\longrightarrow\infty}\|\psi_n - \mathfrak{I}_n\| = 0$.

Proof. The proof will be divided as follows: □

Initialization: select initial $\vartheta_0, \vartheta_1 \in \daleth$, $\rho \in (0,1)$, $\Lambda \geq 0$, $\ell_0 > 0$, and $0 < \beta < 1$.
St. (i). Put $\mathfrak{I}_n$ as:
$$\mathfrak{I}_n = \vartheta_n + \Lambda(\vartheta_n - \vartheta_{n-1}),$$
St. (ii). Calculate:
$$\psi_n = (1 + \ell_n \Upsilon)^{-1}(1 - \ell_n ¥)\mathfrak{I}_n.,$$
If $\mathfrak{I}_n = \psi_n$, discontinue. $\mathfrak{I}_n$ is a solution of (1), otherwise, continue to **St. (iii)**
St. (iii). Calculate:
$$\phi_n = (1-\beta)\mathfrak{I}_n + \beta\psi_n + \beta\ell_n(¥\mathfrak{I}_n - ¥\psi_n),$$
where ℓ_{n+1} is stepsize sequence revised as follows:
$$\ell_{n+1} = \begin{cases} \min\{\ell_n, (\rho\|\mathfrak{I}_n - \psi_n\|)/(\|¥\mathfrak{I}_n - ¥\psi_n\|)\}, & \text{if } ¥\mathfrak{I}_n \neq ¥\psi_n, \\ \ell_n, & \text{else,} \end{cases}$$
St. (iv). Calculate:
$$C_{n+1} = \left\{\eta \in C_n : \|\phi_n - \eta\|^2 \leq \|\vartheta_n - \eta\|^2 + \Lambda^2\|\vartheta_{n-1} - \vartheta_n\|^2 - 2\Lambda\langle\vartheta_n - \eta, \vartheta_{n-1} - \vartheta_n\rangle - \beta\Delta\|\mathfrak{I}_n - \psi_n\|^2\right\},$$
where $\Delta = (2 - \beta - 2\rho(1-\beta)\ell_n/\ell_{n+1} - \beta\rho^2\ell_n^2/\ell_{n+1}^2)$.
St. (v). Compute
$$\vartheta_{n+1} = P_{C_{n+1}}(\vartheta_1), \quad n \geq 1,$$
put $n = n+1$, and return to **St. (i).**

Algorithm 1: Splitting method for the VIP.

Part 1. Demonstrate that $P_{C_{n+1}}\vartheta_1$ is well-defined, for each $\vartheta_1 \in \daleth$, $n \geq 1$, and $\Omega \subset C_{n+1}$. It follows from condition (i) and Lemma 4 that $\mho_{\ell_n} = (I + \ell_n\Upsilon)^{-1}(I - \ell_n ¥)$ is a nonexpansive mapping. Lemma 3 implies that Ω is a closed and convex set, and Lemma 1 clarifies that C_{n+1} is closed and convex, for all $n \geq 1$.

Let $\eta \in \Omega$; we have

$$\|\mathfrak{I}_n - \eta^2\| = \|(\vartheta_n - \eta) - \Lambda(\vartheta_{n-1} - \vartheta_n)\|^2 = \|\vartheta_n - \eta\|^2 - 2\Lambda\langle\vartheta_n - \eta, \vartheta_{n-1} - \vartheta_n\rangle + \Lambda^2\|\vartheta_{n-1} - \vartheta_n\|^2. \tag{27}$$

Since the resolvent $\mho_{\ell_n}$ is firmly a nonexpansive mapping and by Lemma 3, we have

$$\begin{aligned}\langle\psi_n - \eta, \mathfrak{I}_n - \psi_n - \ell_n ¥\mathfrak{I}_n\rangle &= \langle J_\ell^\Upsilon(I - \ell_n ¥)\mathfrak{I}_n - J_\ell^\Upsilon(I - \ell_n ¥)\eta, (I - \ell_n ¥)\mathfrak{I}_n - (I - \ell_n ¥)\eta + (I - \ell_n ¥)\eta - \psi_n\rangle \\ &\geq \|\psi_n - \eta\|^2 + \langle\psi_n - \eta, \eta - \psi_n\rangle - \langle\psi_n - \eta, \ell_n ¥\psi_n\rangle = -\langle\psi_n - \eta, \ell_n ¥\psi_n\rangle.\end{aligned} \tag{28}$$

Hence, by (28), we get

$$\langle\psi_n - \eta, \mathfrak{I}_n - \psi_n - \ell_n(¥\mathfrak{I}_n + ¥\psi_n)\rangle \geq 0, \tag{29}$$

which leads to

$$2\langle\mathfrak{I}_n - \psi_n, \psi_n - \eta\rangle - 2\ell_n\langle ¥\mathfrak{I}_n + ¥\psi_n, \psi_n - \eta\rangle \geq 0. \tag{30}$$

It is obvious that

$$2\langle\mathfrak{I}_n - \psi_n, \psi_n - \eta\rangle = \|\mathfrak{I}_n - \eta\|^2 - \|\mathfrak{I}_n - \psi_n\|^2 - \|\psi_n - \eta\|^2. \tag{31}$$

Applying (31) in (30), we can write

$$\|\psi_n - \eta\|^2 \leq \langle\mathfrak{I}_n - \eta\rangle^2 - \|\mathfrak{I}_n - \psi_n\|^2 - 2\ell_n\langle ¥\mathfrak{I}_n - ¥\psi_n, \psi_n - \eta\rangle. \tag{32}$$

Now, from definition ϕ_n, we have

$$\begin{aligned}\|\phi_n - \eta\|^2 &= \|(1-\beta)\mathfrak{I}_n + \beta\psi_n + \beta\ell_n(¥\mathfrak{I}_n - ¥\psi_n) - \eta\|^2 = \|(1-\beta)(\mathfrak{I}_n - \eta) + \beta(\psi_n - \eta) + \beta\ell_n(¥\mathfrak{I}_n - ¥\psi_n)\|^2 \\ &= (1-\beta)^2\|\mathfrak{I}_n - \eta\|^2 + \beta^2\|\psi_n - \eta\| + \beta^2\ell_n^2\|¥\mathfrak{I}_n - ¥\psi_n\|^2 + 2\beta(1-\beta)\langle\mathfrak{I}_n - \eta, \psi_n - \eta\rangle \\ &\quad + 2\beta\ell_n(1-\beta)\langle\mathfrak{I}_n - \eta, ¥\mathfrak{I}_n - ¥\psi_n\rangle + 2\beta^2\ell_n\langle\psi_n - \eta, ¥\mathfrak{I}_n - ¥\psi_n\rangle.\end{aligned} \tag{33}$$

From equation (15), one can write

$$2\langle \mathfrak{I}_n - \eta, \psi_n - \eta\rangle = \|\mathfrak{I}_n - \eta\|^2 - \|\mathfrak{I}_n - \psi_n\|^2 + \|\psi_n - \eta\|^2. \tag{34}$$

Applying (34) in (33), we get

$$\begin{aligned}\|\phi_n - \eta\|^2 = {} & (1-\beta)\|\mathfrak{I}_n - \eta\|^2 + \beta\|\psi_n - \eta\|^2 - \beta(1-\beta)\|\psi_n - \mathfrak{I}_n\|^2 + \beta^2\ell_n^2\|¥\mathfrak{I}_n - ¥\psi_n\|^2 + 2\beta\ell_n(1-\beta)\langle \mathfrak{I}_n - \eta, ¥\mathfrak{I}_n - ¥\psi_n\rangle \\ & + 2\beta^2\ell_n\langle \psi_n - \eta, ¥\mathfrak{I}_n - ¥\psi_n\rangle.\end{aligned} \tag{35}$$

It follows from (32), (35), and (26) that

$$\begin{aligned}\|\phi_n - \eta\|^2 &\le (1-\beta)\|\mathfrak{I}_n - \eta\|^2 + \beta\left[\|\mathfrak{I}_n - \eta\|^2 - \|\mathfrak{I}_n - \psi_n\|^2 - 2\ell_n\langle ¥\mathfrak{I}_n - ¥\psi_n, \psi_n - \eta\rangle\right] - \beta(1-\beta)\|\psi_n - \mathfrak{I}_n\|^2 \\ &\quad + \beta^2\ell_n^2\|¥\mathfrak{I}_n - ¥\psi_n\|^2 + 2\beta\ell_n(1-\beta)\langle \mathfrak{I}_n - \eta, ¥\mathfrak{I}_n - ¥\psi_n\rangle + 2\beta^2\ell_n\langle \psi_n - \eta, ¥\mathfrak{I}_n - ¥\psi_n\rangle \\ &\le \|\mathfrak{I}_n - \eta\|^2 - \beta(2-\beta)\|\mathfrak{I}_n - \psi_n\|^2 - 2\beta\ell_n\langle ¥\mathfrak{I}_n - ¥\psi_n, \psi_n - \eta\rangle + \beta^2\ell_n^2\|¥\mathfrak{I}_n - ¥\psi_n\|^2 \\ &\quad + 2\beta\ell_n(1-\beta)\langle \mathfrak{I}_n - \eta, ¥\mathfrak{I}_n - ¥\psi_n\rangle + 2\beta\ell_n\langle \psi_n - \eta, ¥\mathfrak{I}_n - ¥\psi_n\rangle \\ &\le \|\mathfrak{I}_n - \eta\|^2 - \beta(2-\beta)\|\mathfrak{I}_n - \psi_n\|^2 + \beta^2\ell_n^2\|¥\mathfrak{I}_n - ¥\psi_n\|^2 + 2\beta\ell_n(1-\beta)\langle \mathfrak{I}_n - \psi_n, ¥\mathfrak{I}_n - ¥\psi_n\rangle \\ &\le \|\mathfrak{I}_n - \eta\|^2 - \beta(2-\beta)\|\mathfrak{I}_n - \psi_n\|^2 + \beta^2\ell_n^2\frac{\rho^2}{\ell_{n+1}^2}\|\mathfrak{I}_n - \psi_n\|^2 + 2\beta\ell_n(1-\beta)\frac{\rho}{\ell_{n+1}}\|\mathfrak{I}_n - \psi_n\|^2 \\ &= \|\mathfrak{I}_n - \eta\|^2 - \beta\left[2 - \beta - 2\rho(1-\beta)\frac{\ell_n}{\ell_{n+1}} - \beta\rho^2\frac{\ell_n^2}{\ell_{n+1}^2}\right]\|\mathfrak{I}_n - \psi_n\|^2 = \|\mathfrak{I}_n - \eta\|^2 - \beta\Delta_n\|\mathfrak{I}_n - \psi_n\|^2.\end{aligned} \tag{36}$$

Applying (27) in (36), we have

$$\begin{aligned}\|\phi_n - \eta\|^2 \le {} & \|\vartheta_n - \eta\|^2 + \Lambda^2\|\vartheta_{n-1} - \vartheta_n\|^2 \\ & - 2\Lambda\langle \vartheta_n - \eta, \vartheta_{n-1} - \vartheta_n\rangle - \beta\Delta_n\|\mathfrak{I}_n - \psi_n\|^2.\end{aligned} \tag{37}$$

It is clear that $\Omega \subset C_1 = \top$. Assume that $\Omega \subset C_n$ for some $n \ge 1$. Then, $\eta \in C_n$ and by (37), we have for all $n \ge 1$, $\eta \in C_{n+1}$. Thus, $\Omega \subset C_{n+1}$ for all $n \ge 1$, i.e., $P_{C_{n+1}}\vartheta_1$ is well-defined and bounded.

Part 2. Illustrate that $\{\vartheta_n\}$ is bounded. Since $\Omega \ne \varnothing$ and closed and convex subset of $\top$, there is a unique $u \in \Omega$ such that $u = P_\Omega\vartheta_1$. This leads to $\vartheta_n = P_{C_n}\vartheta_1$, $C_n \subset C_{n+1}$, and $\vartheta_{n+1} \in C_n$ for all $n \ge 1$, and we have

$$\|\vartheta_n - \vartheta_1\| \le \|\vartheta_{n+1} - \vartheta_1\|. \tag{38}$$

Furthermore, as $\Omega \subset C_n$, for all $n \ge 1$, we obtain

$$\|\vartheta_n - \vartheta_1\| \le \|u - \vartheta_1\|. \tag{39}$$

It follows by (38) and (39) that $\lim_{n\longrightarrow\infty}\|\vartheta_n - \vartheta_1\|$ exists. Hence, $\{\vartheta_n\}$ is bounded.

Part 3. Fulfillment of $\lim_{n\longrightarrow\infty}\vartheta_n = \tau$. By the definition of C_n, for $m > n$, we observe that $\vartheta_m = P_{C_m}\vartheta_1 \in C_m \subset C_n$. From Lemma 2, we have

$$\|\vartheta_m - \vartheta_n\|^2 \le \|\vartheta_m - \vartheta_1\|^2 - \|\vartheta_n - \vartheta_1\|^2. \tag{40}$$

By Part 2, we conclude that $\lim_{n,m\longrightarrow\infty}\|\vartheta_m - \vartheta_n\|^2 = 0$. Thus, $\{\vartheta_n\}$ is a Cauchy sequence. Hence, $\lim_{n\longrightarrow\infty}\vartheta_n = \tau$. Additionally, we get

$$\lim_{n\longrightarrow\infty}\|\vartheta_{n+1} - \vartheta_n\| = 0. \tag{41}$$

Part 4. Prove that $\tau \in \Omega$. It follows from (41) that

$$\|\mathfrak{I}_n - \vartheta_n\| = \Lambda\|\vartheta_n - \vartheta_{n-1}\| \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{42}$$

Also, by (42) and condition (ii), we can write

$$\|\psi_n - \vartheta_n\| \le \|\psi_n - \mathfrak{I}_n\| + \|\mathfrak{I}_n - \vartheta_n\| \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{43}$$

From triangle inequality on the norm and (42) and (43), we obtain

$$\|\mathfrak{J}_n - \psi_n\| \le \|\mathfrak{J}_n - \vartheta_n\| + \|\psi_n - \vartheta_n\| \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{44}$$

Replacing η with ϑ_n in (36) and using (41) and (44), we have

$$\|\phi_n - \vartheta_n\|^2 \le \Lambda^2 \|\vartheta_{n-1} - \vartheta_n\|^2 - \beta\Delta_n \|\mathfrak{J}_n - \psi_n\|^2 \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{45}$$

Applying (41), (42), and (45), we can write

$$\begin{aligned} \|\vartheta_{n+1} - \mathfrak{J}_n\| &\le \|\vartheta_{n+1} - \vartheta_n\| + \|\mathfrak{J}_n - \vartheta_n\| \longrightarrow 0 \text{ as } n \longrightarrow \infty, \\ \|\vartheta_{n+1} - \phi_n\| &\le \|\vartheta_{n+1} - \vartheta_n\| + \|\phi_n - \vartheta_n\| \longrightarrow 0 \text{ as } n \longrightarrow \infty, \\ \|\phi_n - \mathfrak{J}_n\| &\le \|\phi_n - \vartheta_n\| + \|\mathfrak{J}_n - \vartheta_n\| \longrightarrow 0 \text{ as } n \longrightarrow \infty. \end{aligned} \tag{46}$$

It follows from (44) that

$$\lim_{n\longrightarrow\infty} \|\mathcal{O}_{\ell_n}\mathfrak{J}_n - \mathfrak{J}_n\| = \lim_{n\longrightarrow\infty} \|\psi_n - \mathfrak{J}_n\| = 0. \tag{47}$$

Since $\liminf_{n\longrightarrow\infty} \ell_n > 0$, there is $\varepsilon > 0$ such that $\ell_n \ge \varepsilon$ and $\varepsilon \in (0, 2\Lambda)$ for all $n \ge 1$. Then, by Lemma 3 (ii) and (47), we get

$$\|\mathcal{O}_\varepsilon \mathfrak{J}_n - \mathfrak{J}_n\| \le 2\|\mathcal{O}_{\ell_n}\mathfrak{J}_n - \mathfrak{J}_n\| \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{48}$$

From (45) and (46), since $\vartheta_n \longrightarrow \tau$ as $n \longrightarrow \infty$, we have also $\mathfrak{J}_n \longrightarrow \tau$ as $n \longrightarrow \infty$. Since $\mathcal{O}_\varepsilon$ is a nonexpansive and continuous mapping, from (47), we conclude that $\tau \in \Omega$.

Part 5. Show that $\tau = P_\Omega(\vartheta_1)$. Since $\vartheta_n = P_{C_n}\vartheta_1$ and $\Omega \subset C_n$, we can get

$$\langle \vartheta_1 - \vartheta_n, \vartheta_n - \eta \rangle \ge 0, \quad \forall \eta \in \Omega. \tag{49}$$

Setting $n \longrightarrow \infty$ in (49), we have

$$\langle \vartheta_1 - \tau, \tau - \eta \rangle \ge 0, \quad \forall \eta \in \Omega. \tag{50}$$

This shows that $\tau = P_\Omega(\vartheta_1)$. This finishes the proof.

4. Solve a Minimization Problem

As an application of our theorem, we solve the following constrained convex minimization problem:

$$\min_{\vartheta \in C} \beth(\vartheta), \tag{51}$$

where $\beth: \daleth \longrightarrow \mathbb{R}$ is a convex function. We suppose that the function $\beth$ is differentiable such that $\nabla\beth$ is an Λ–ism operator.

It is easy to see that problem (51) is equivalent to the following problem:

$$\min_{\vartheta \in \daleth} [\beth(\vartheta) + \wp_C(\vartheta)], \tag{52}$$

where $\wp_C$ is the indicator function of C. Thus, this problem becomes the problem of finding an element $\vartheta^* \in \daleth$ such that

$$\nabla\beth(\vartheta^*) + \partial\wp_C(\vartheta^*) \ni 0, \tag{53}$$

where $\partial\wp_C$ is the subdifferential of $\wp_C$. We know that $\partial\wp_C$ is a maximal monotone operator, and $(I + m\,\partial\wp_C)^{-1} = P_C$ for all $m > 0$.

For solving problem (51), we state the theorem in the following, which is similar to Theorem 1.

Theorem 2. *Let the sequence $\{\ell_n\}$ be bounded below by $\min\{(\rho/L), \ell_0\}$, where $\rho \in (0,1)$ and $\ell_0 > 0$. Given a parameter $\Lambda \ge 0$ such that $0 < \inf\{\ell_n\} \le sup_n\{\ell_n\} < 2\Lambda$. Let $\{\vartheta_n\}$ be the sequence in $\daleth$ which is defined by $\vartheta_0, \vartheta_1 \in \daleth$, $C_1 = \daleth$, $0 < \beta < 1$, and*

$$\begin{cases} \mathfrak{J}_n = \vartheta_n + \Lambda(\vartheta_n - \vartheta_{n-1}), \\ \psi_n = P_C(\mathfrak{J}_n - \ell_n \nabla\beth\mathfrak{J}_n), \\ \phi_n = (1-\beta)\mathfrak{J}_n + \beta\psi_n + \beta\ell_n(¥\mathfrak{J}_n - ¥\psi_n), \\ \text{where,} \quad \ell_{n+1} = \begin{cases} \min\left\{\ell_n, \dfrac{\rho\|\mathfrak{J}_n - \psi_n\|}{\|¥\mathfrak{J}_n - ¥\psi_n\|}\right\}, & \text{if } ¥\mathfrak{J}_n \ne ¥\psi_n, \\ \ell_n, & \text{else,} \end{cases} \\ C_{n+1} = \left\{ \begin{array}{c} \eta \in \daleth: \|\phi_n - \eta\|^2 \le \|\vartheta_n - \eta\|^2 + \Lambda^2\|\vartheta_{n-1} - \vartheta_n\|^2 \\ -2\Lambda\langle \vartheta_n - \eta, \vartheta_{n-1} - \vartheta_n\rangle - \beta\Delta\|\mathfrak{J}_n - \psi_n\|^2 \end{array} \right\}, \\ \vartheta_{n+1} = P_{C_{n+1}}(\vartheta_1),\ n \ge 1, \end{cases} \tag{54}$$

where ¥: $\daleth \longrightarrow \daleth$ is Λ–ism on a RHS $\daleth$, $\Upsilon: \daleth \longrightarrow 2^{\daleth}$ is a maximally monotone operator, and $\Delta = (2 - \beta - 2\rho(1-\beta)\ell_n/\ell_{n+1} - \beta\rho^2\ell_n^2/\ell_{n+1}^2)$. If $\Omega \ne \varnothing$, then the sequence $\{\vartheta_n\}$ converges strongly to $\tau = P_\Omega(\vartheta_1)$, provided that $lim_{n\longrightarrow\infty}\|\psi_n - \mathfrak{J}_n\| = 0$.

5. Solve a Split Feasibility Problem

In this section, we investigated the application of our proposed methods to the split convex feasibility problem (SCFP). Let $T: \daleth_1 \longrightarrow \daleth_2$ be a bounded linear operator and T^* its adjoint defined on the two RHSs $\daleth_1$ and $\daleth_2$. Assume that $\mathscr{C} \subset \daleth_1$ and $\mathscr{Q} \subset \daleth_2$ are nonempty CCSs. The SCFP [54] take the shape as follows:

$$\text{create a point } \vartheta \in \mathscr{C} \text{ so that } T(\vartheta) \in \mathscr{Q}. \tag{55}$$

In a HS, SFP was initiated by Censor and Elfving [54], and they used a multidistance approach to find an adaptive approach for resolving it. Many of the problems that emerge from state retrieval and restoration of medical image can be formulated as SVFP [55, 56]. SFP is also used in a variety of disciplines such as dynamic emission tomographic image

reconstruction, image restoration, and radiation therapy treatment planning [57–59]. Let us consider

$$¥(\vartheta) := \nabla\left(\frac{1}{2}\left\|T\vartheta - P_Q(T\vartheta)\right\|^2\right) = T^*\left(I - P_Q\right)T\vartheta \tag{56}$$

for the metric projection P_Q on to Q, the gradient ∇, and $\Upsilon = \partial i_{\mathscr{C}}$. Due to the above construction, problem (55) has an inclusion format as described in (1). It can be seen that ¥ is Lipschitz continuous with constant $L = \|T\|^2$, and Υ is maximal monotone, see, e.g., [60].

Let $\mathscr{C}$ be a nonempty CCS of a RHS $\daleth$, and a normal cone of $\mathscr{C}$ at $\vartheta \in \mathscr{C}$ is defined by

$$N_{\mathscr{C}}(\vartheta) = \{z \in \daleth\colon \langle z, y - \vartheta\rangle \le 0, \forall y \in \mathscr{C}\}. \tag{57}$$

Suppose $g\colon \daleth \longrightarrow (-\infty, +\infty)$ is a proper, lower semicontinuous, and convex function. For each $\vartheta \in \daleth$, the subdifferential ∂g of g is given by

$$\partial g(\vartheta) = \{z \in \daleth\colon g(y) - g(\vartheta) \ge \langle z, y - \vartheta\rangle, \quad \forall y \in \mathscr{C}\}. \tag{58}$$

For any nonempty CCS $\mathscr{C}$ of $\daleth$, the indicator function $i_{\mathscr{C}}$ of $\mathscr{C}$ is defined by

$$i_{\mathscr{C}}(\vartheta) = \begin{cases} 0, & \text{if } \vartheta \in \mathscr{C} \\ \infty, & \text{otherwise.} \end{cases} \tag{59}$$

It is obvious that the indicator function $i_{\mathscr{C}}$ is proper, convex, and lower semicontinuous on $\daleth$. A subdifferential $\partial i_{\mathscr{C}}$ of $i_{\mathscr{C}}$ is a maximal monotone operator, and

$$\begin{aligned}\partial i_{\mathscr{C}}(\vartheta) &= \{z \in \daleth\colon i_{\mathscr{C}}(y) - i_{\mathscr{C}}(\vartheta) \ge \langle z, y - \vartheta\rangle, \forall y \in \mathscr{C}\} \\ &= \{z \in \daleth\colon \langle z, y - \vartheta\rangle \le 0, \forall y \in \mathscr{C}\} = N_{\mathscr{C}}(\vartheta).\end{aligned} \tag{60}$$

For each $\vartheta \in \daleth$, now we define the resolvent of an indicator function $\partial i_{\mathscr{C}}$ for each $\lambda > 0$ in the following manner:

$$J_\lambda^{\partial i_{\mathscr{C}}} = \left(\mathrm{Id} + \lambda\, \partial i_{\mathscr{C}}\right)^{-1}. \tag{61}$$

Hence, we can observe that

$$\begin{aligned} y = J_\lambda^{\partial i_{\mathscr{C}}}(\vartheta) &\Longleftrightarrow \vartheta \in \left(y + \lambda\, \partial i_{\mathscr{C}}(y)\right)^{-1} \Longleftrightarrow \vartheta - y \in \lambda\, \partial i_{\mathscr{C}}(y) \\ &\Longleftrightarrow y = P_{\mathscr{C}}(\vartheta). \end{aligned} \tag{62}$$

Now, on the basis of the above, Algorithm 1 may be reduced to the following scheme.

Theorem 3. *Let $\{\vartheta_n\}$ be a sequence generated by the following scheme: choose $\vartheta_{-1}, \vartheta_0 \in \mathscr{C}$, $\rho \in (0,1)$, $\Lambda \ge 0$, $\ell_0 > 0$, and $0 < \beta < 1$.*

St. (i): compute $\mathfrak{I}_n$ in the following way:

$$\mathfrak{I}_n = \vartheta_n + \Lambda(\vartheta_n - \vartheta_{n-1}). \tag{63}$$

St. (ii): calculate

$$\psi_n = P_{\mathscr{C}}\left[\mathfrak{I}_n - \ell_n T^*\left(I - P_Q\right)T\mathfrak{I}_n\right]. \tag{64}$$

If $\mathfrak{I}_n = \psi_n$, stop, and $\mathfrak{I}_n$ is a solution of problem (55); otherwise, continue to St. (iii).

St. (iii): calculate

$$\phi_n = (1-\beta)\mathfrak{I}_n + \beta\psi_n + \beta\ell_n\left[T^*\left(I - P_Q\right)T\mathfrak{I}_n - T^*\left(I - P_Q\right)T\psi_n\right], \tag{65}$$

where ℓ_{n+1} is the stepsize sequence revised in the following way:

$$\ell_{n+1} = \begin{cases} \min\left\{\ell_n, \dfrac{\rho\|\mathfrak{I}_n - \psi_n\|}{\left[T^*\left(I - P_Q\right)T\mathfrak{I}_n - T^*\left(I - P_Q\right)T\psi_n\right]}\right\}, & \text{if } T^*\left(I - P_Q\right)T\mathfrak{I}_n \ne T^*\left(I - P_Q\right)T\psi_n, \\ \ell_n, & \text{otherwise.} \end{cases} \tag{66}$$

St. (iv): calculate

$$C_{n+1} = \left\{\eta \in \daleth\colon \|\phi_n - \eta\|^2 \le \|\vartheta_n - \eta\|^2 + \Lambda^2\|\vartheta_{n-1} - \vartheta_n\|^2 - 2\Lambda\langle\vartheta_n - \eta, \vartheta_{n-1} - \vartheta_n\rangle - \beta\Delta\|\mathfrak{I}_n - \psi_n\|^2\right\}, \tag{67}$$

where $\Delta = (2 - \beta - 2\rho(1-\beta)\ell_n/\ell_{n+1} - \beta\rho^2\ell_n^2/\ell_{n+1}^2)$.

St. (v): compute

$$\vartheta_{n+1} = P_{C_{n+1}}(\vartheta_1), \quad n \ge 1. \tag{68}$$

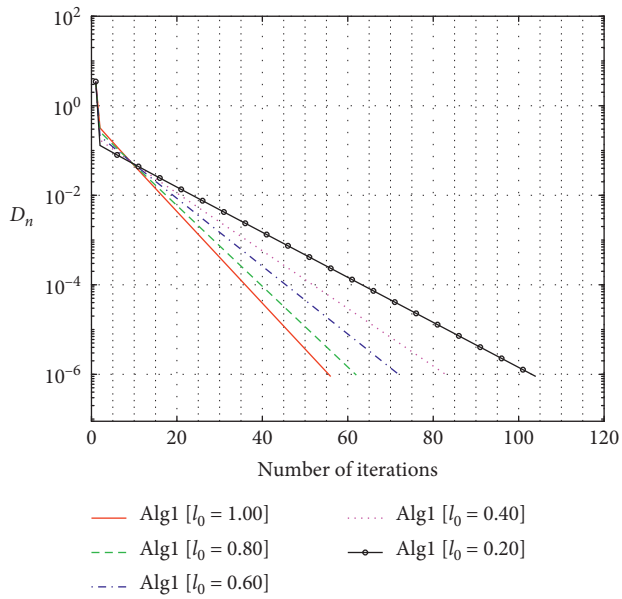


Figure 1: Numerical conduct of Alg1 by choosing different values of ℓ_0.

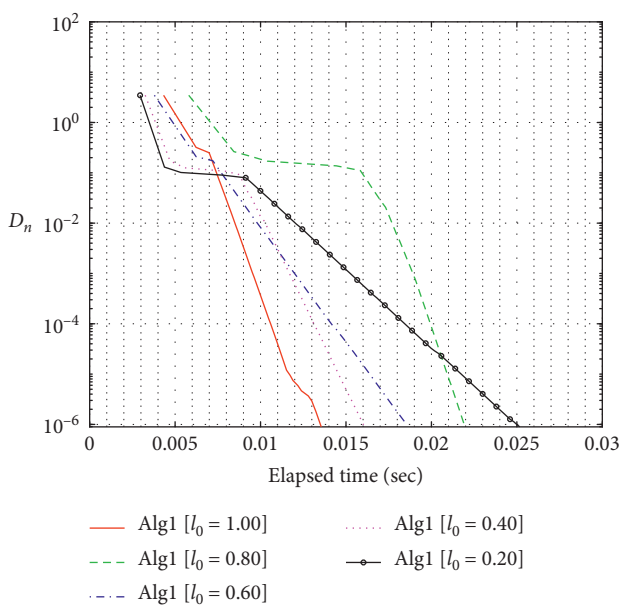


Figure 2: Numerical conduct of Alg1 by choosing different values of ℓ_0.

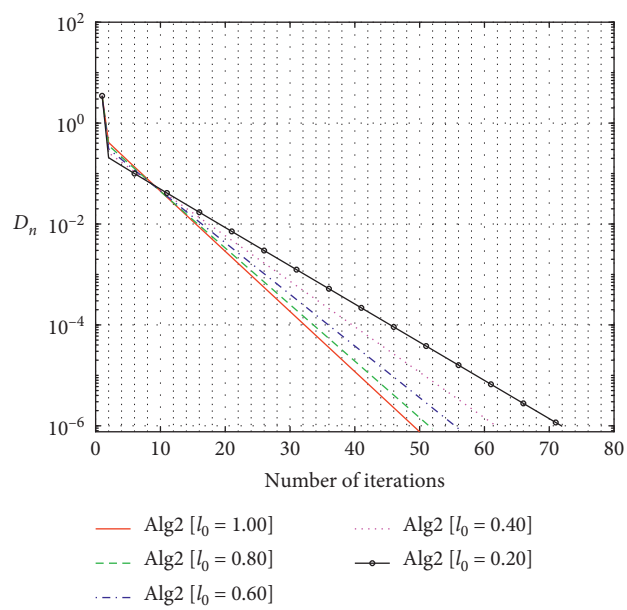


Figure 3: Numerical conduct of Alg2 by choosing different values of ℓ_0.

Put $n = n + 1$, and return to St. (i). If the solution set Γ_{SFP} is nonempty, then the sequence $\{\vartheta_n\}$ converges weakly to an element of $\Gamma_{(SFP)}$.

6. Numerical Discussion

This part is devoted to present a numerical solution to a SCFP in an infinite HS, which is a special inclusion problem as explained in Section 5. The problem setting is taken from [61]. We provide the comparison of Algorithm 1 (Alg1) in [45] and our proposed Algorithm 1 (Alg2).

Example 1. Let $\daleth_1 = \daleth_2 = L_2([0, 2\pi])$ be two HSs with an inner product

$$\langle \vartheta, y \rangle := \int_0^{2\pi} \vartheta(t) y(t) \mathrm{d}t, \quad \forall \vartheta, y \in L_2([0, 2\pi]), \tag{69}$$

and the induced norm defined by

$$\|\vartheta\| := \sqrt{\int_0^{2\pi} |\vartheta(t)|^2 \mathrm{d}t}, \quad \forall \vartheta \in L_2([0, 2\pi]). \tag{70}$$

Next, consider the feasible set $\mathscr{C} \subset \daleth_1$ as

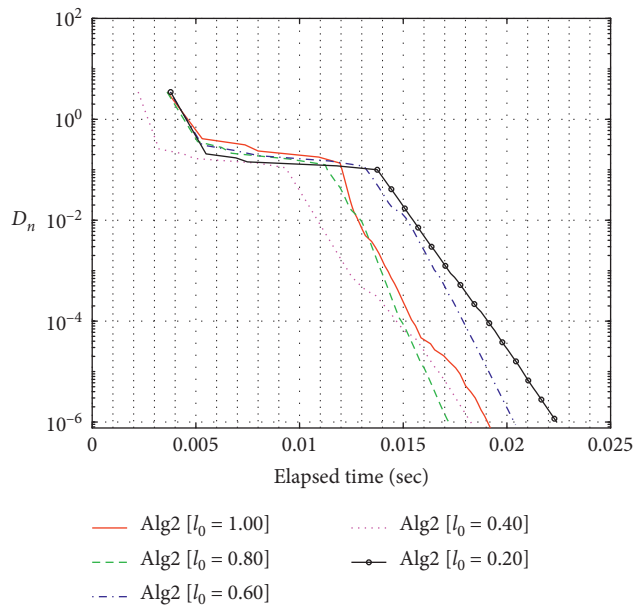


Figure 4: Numerical conduct of Alg2 by choosing different values of ℓ_0.

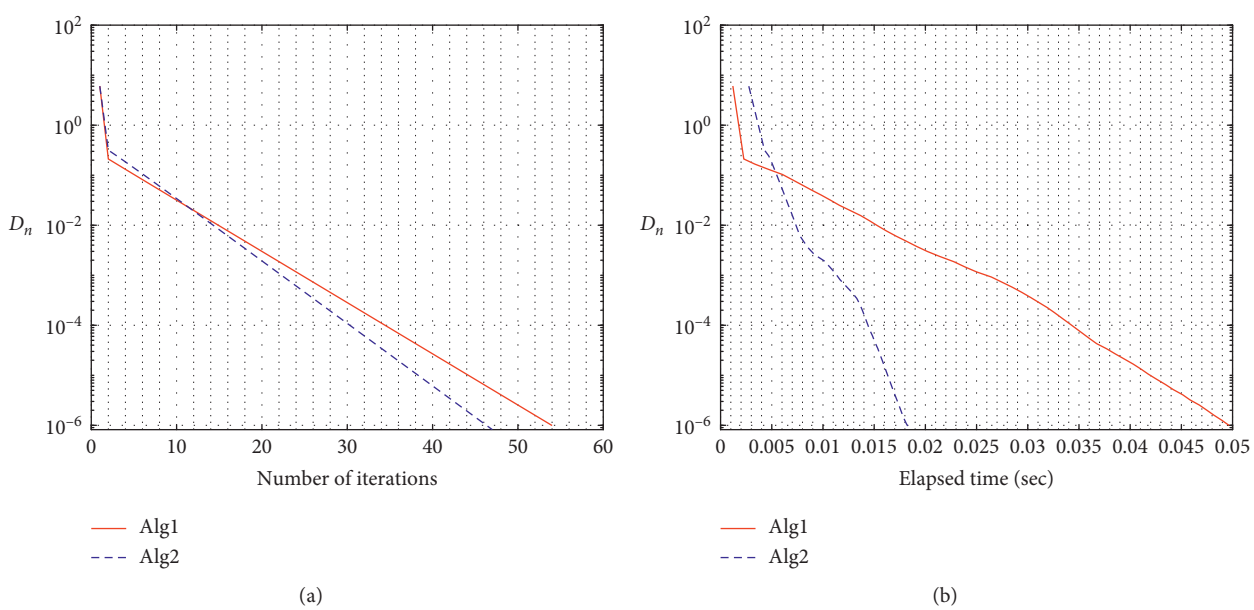


Figure 5: Numerical comparison of Alg2 with Alg.1 by assuming values of $\vartheta_{-1} = \vartheta_0 = t$.

$$\mathscr{C} = \left\{ \vartheta \in \daleth_1 : \int_0^{2\pi} \vartheta(t)\mathrm{d}t \leq 1 \right\}, \tag{71}$$

and $\mathscr{Q} \subset \daleth_2$ is

$$\mathscr{Q} = \left\{ \vartheta \in \daleth_2 : \int_0^{2\pi} |\vartheta(t) - \sin(t)|^2 \mathrm{d}t \leq 16 \right\}. \tag{72}$$

Consider the mapping $T: \daleth_1 \longrightarrow \daleth_2$ such that $(T\vartheta)(s) = \vartheta(s)$, $\vartheta \in \daleth_1$. Then, $(T^*\vartheta)(s) = \vartheta(s)$, and $\|T\| = 1$. So, we shall solve the following problem:

$$\text{create } \vartheta^* \in \mathscr{C} \text{ so that } T(\vartheta^*) \in \mathscr{Q}. \tag{73}$$

We can also observe that since $(T\vartheta)(s) = \vartheta(s)$, $\vartheta \in \daleth_1$, the above problem is actually a CFP of the form

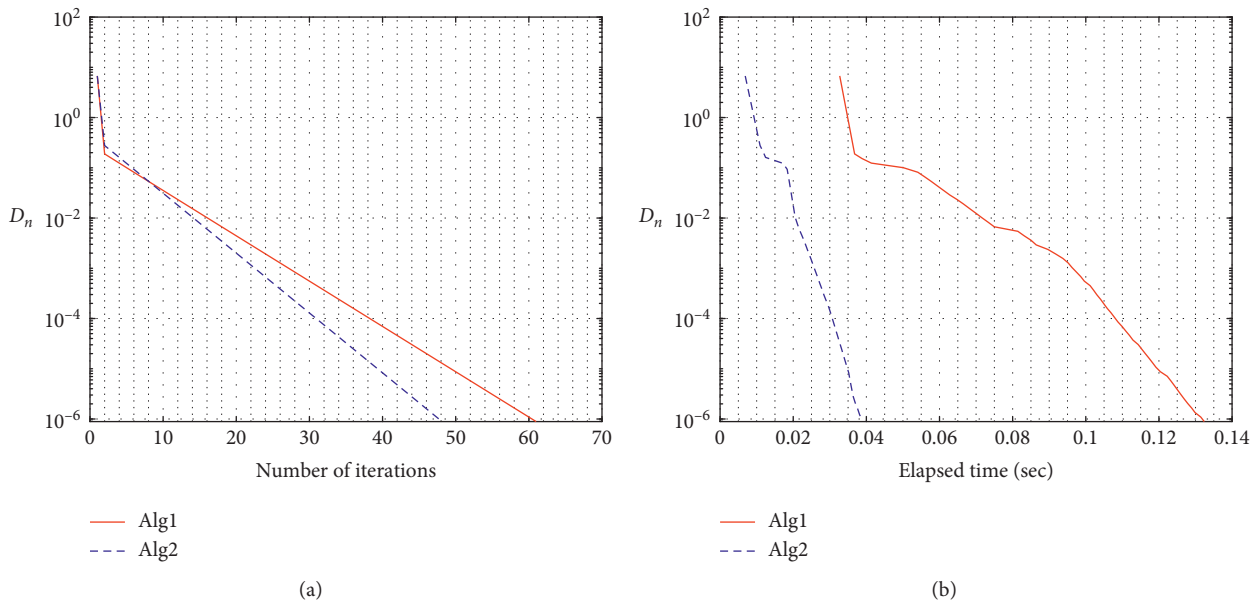


Figure 6: Numerical comparison of Alg2 with Alg.1 by assuming values of $\vartheta_{-1} = \vartheta_0 = t^2/5$.

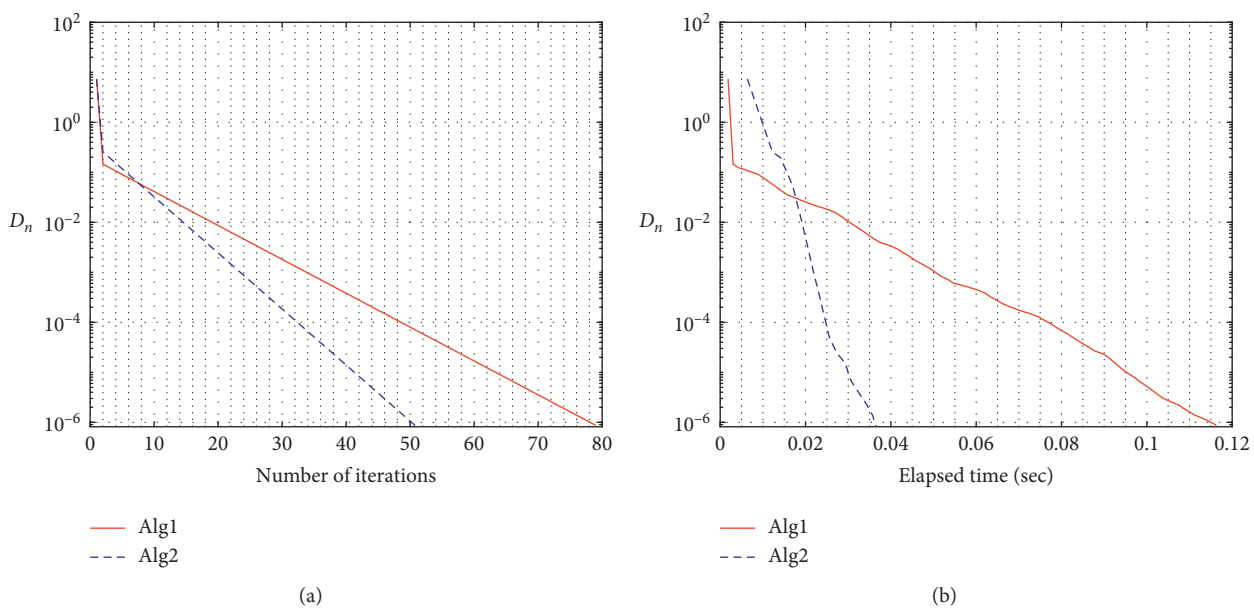


Figure 7: Numerical comparison of Alg2 with Alg1 by assuming values of $\vartheta_{-1} = \vartheta_0 = 2e^t t^5$.

$$\text{create } \vartheta^* \in \mathscr{C} \cap \mathscr{Q}. \tag{74}$$

Figures 1–9 and Tables 1 and 2 show the numerical results by assuming $D_n = \|\vartheta_n - \vartheta_{n_1}\| \leq 10^{-6}$.

Remark 1. It is well known that the success of any iterative method depends on two main things: first, the number of iterations: when the number of iterations is small, the method is successful in saving effort. Second, time factor: the method that needs less time in implementation is excellent than its counterpart, which needs a lot of time and is considered successful in saving time. So, from figures and tables, we observe that our algorithm needs fewer iterations and less time than Algorithm 1 [45]. This illustrates that our method is successful in speeding up Algorithm 1 [45] and solving problem (55). Also, the performance of our algorithm is good because it saves time and effort in studding the convergence rate.

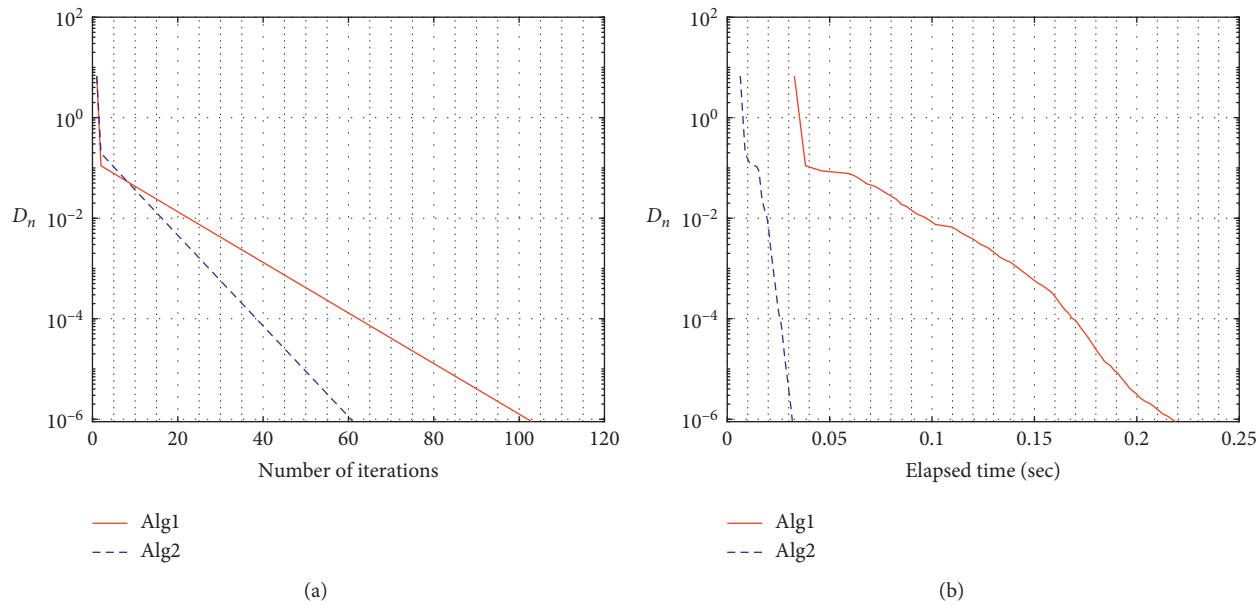


Figure 8: Numerical comparison of Alg2 with Alg1 by assuming values of $\vartheta_{-1} = \vartheta_0 = e^t \sin(t)$.

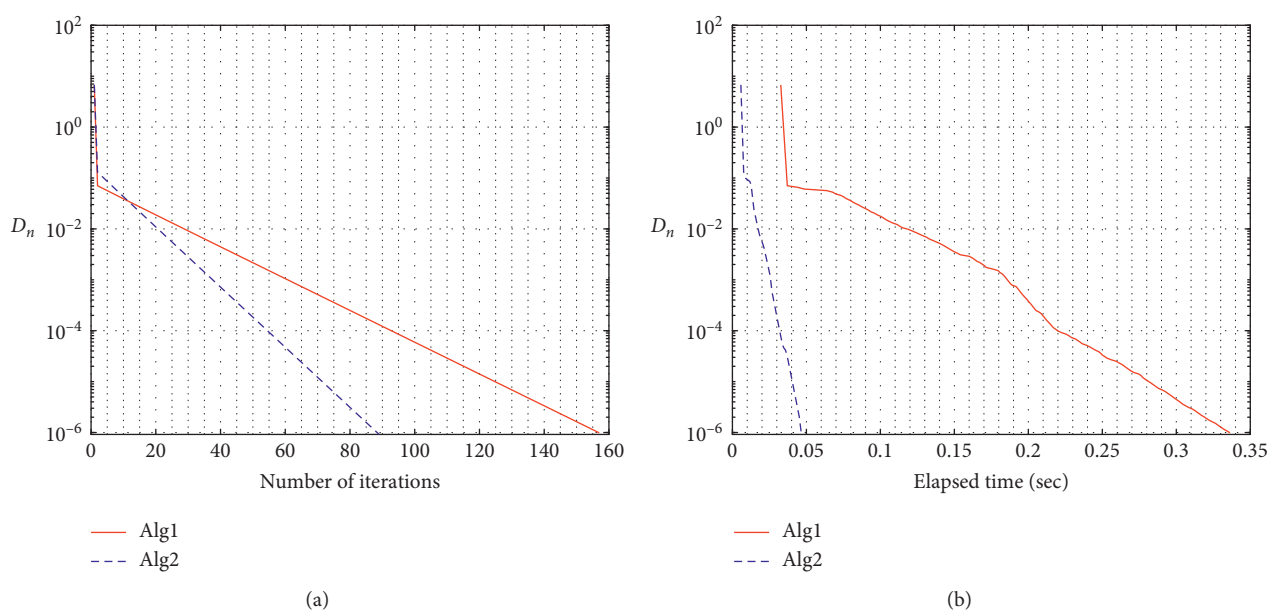


Figure 9: Numerical comparison of Alg2 with Alg1 by assuming values of $\vartheta_{-1} = \vartheta_0 = (t^2 - e^t)\cos(t)$.

Table 1: Numerical comparison of Alg2 with Alg1 by assuming different values of ℓ_0.

				Number of iterations		Execution time in seconds	
$\vartheta_{-1} = \vartheta_0$	ρ	Λ	ℓ_0	Alg1	Alg2	Alg1	Alg2
$1/5 \exp\,(t/2)^{5/4}$	0.27	0.50	1.00	56	50	0.0136	0.0190
$1/5 \exp\,(t/2)^{5/4}$	0.27	0.50	0.80	62	52	0.0219	0.0150
$1/5 \exp\,(t/2)^{5/4}$	0.27	0.50	0.60	72	56	0.0186	0.0205
$1/5 \exp\,(t/2)^{5/4}$	0.27	0.50	0.40	83	62	0.0160	0.0183
$1/5 \exp\,(t/2)^{5/4}$	0.27	0.50	0.20	104	72	0.0252	0.0225

Table 2: Numerical comparison of Alg2 with Alg1.

				Number of iterations		Execution time in seconds	
$\vartheta_{-1} = \vartheta_0$	ρ	Λ	ℓ_0	Alg1	Alg2	Alg1	Alg2
t	0.33	0.35	0.50	54	47	0.0497	0.0184
$t^2/5$	0.33	0.35	0.50	61	48	0.1325	0.0390
$2e^t t^5$	0.33	0.35	0.50	71	51	0.1166	0.0366
$e^t \sin(t)$	0.3	0.35	0.50	103	61	0.2193	0.0318
$(t^2 - e^t)\cos(t)$	0.33	0.35	0.50	157	89	0.3363	0.0467

Data Availability

Data sharing is not applicable to this article as no datasets are generated or analyzed during the current study.

Conflicts of Interest

The authors declare that they have no conflicts of interest concerning the publication of this article.

Authors' Contributions

All authors contributed equally and significantly to writing this article.

Acknowledgments

The authors are grateful to the Spanish Government and the European Commission for Grants IT1207-19 and RTI2018-094336-BI00 (MCIU/AEI/FEDER, UE). This work was supported in part by the Basque Government under Grant IT1207-19.

References

[1] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[2] C. Byrne, "A unified treatment of some iterative algorithms in signal processing and image reconstruction," *Inverse Problems*, vol. 20, no. 1, p. 103, 2003.

[3] C. Byrne, "Iterative oblique projection onto convex sets and the split feasibility problem," *Inverse Problems*, vol. 18, no. 2, p. 441, 2002.

[4] A. Hanjing and S. Suantai, "A fast image restoration algorithm based on a fixed point and optimization method," *Mathematics*, vol. 8, no. 3, p. 378, 2020.

[5] D. V. Thong and P. Cholamjiak, "Strong convergence of a forward–backward splitting method with a new step size for solving monotone inclusions," *Journal of Computational and Applied Mathematics*, vol. 38, p. 94, 2019.

[6] P. Marcotte, "Application of khobotov's algorithm to variational inequalities and network equilibrium problems," *INFOR: Information Systems and Operational Research*, vol. 29, no. 4, pp. 258–270, 1991.

[7] A. Gibali and D. V. Thong, "Tseng type methods for solving inclusion problems and its applications," *Calcolo*, vol. 55, p. 49, 2018.

[8] E. N. Khobotov, "Modification of the extra-gradient method for solving variational inequalities and certain optimization problems," *USSR Computational Mathematics and Mathematical Physics*, vol. 27, no. 5, pp. 120–127, 1987.

[9] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, Cambridge, MA, USA, 1980.

[10] R. Trémolières, J. L. Lions, and R. Glowinski, *Numerical Analysis of Variational Inequalities*, North Holland, Amsterdam, The Netherlands, 2011.

[11] C. Baiocchi, "Variational and quasivariational inequalities," in *Applications to Free-Boundary Problems*Springer, Basel, Switzerland, 1984.

[12] I. Konnov, *Combined Relaxation Methods for Variational Inequalities*, Springer science & Business Media, Berlin, Germany, 2001.

[13] W. Kumam, H. Piri, and P. Kumam, "Solutions of system of equilibrium and variational inequality problems on fixed points of infinite family of nonexpansive mappings," *Applied Mathematics and Computation*, vol. 248, pp. 441–455, 2014.

[14] T. Chamnarnpan, S. Phiangsungnoen, and P. Kumam, "A new hybrid extragradient algorithm for solving the equilibrium and variational inequality problems," *Afrika Matematika*, vol. 26, no. 1-2, pp. 87–98, 2015.

[15] J. Deepho, W. Kumam, and P. Kumam, "A new hybrid projection algorithm for solving the split generalized equilibrium problems and the system of variational inequality problems," *Journal of Mathematical Modelling and Algorithms in Operations Research*, vol. 13, no. 4, pp. 405–423, 2014.

[16] H. A. Hammad, H. ur Rehman, and M. De La Sen, "Advanced algorithms and common solutions to variational inequalities," *Symmetry*, vol. 12, no. 7, p. 1198, 2020.

[17] G. López, V. Martín-Márquez, F. Wang, and H.-K. Xu, "Forward-Backward splitting methods for accretive operators in Banach spaces," *Abstract and Applied Analysis*, vol. 2012, Article ID 109236, 25 pages, 2012.

[18] G. B. Passty, "Ergodic convergence to a zero of the sum of monotone operators in Hilbert space," *Journal of Mathematical Analysis and Applications*, vol. 72, no. 2, pp. 383–390, 1979.

[19] D. W. Peaceman and H. H. Rachford Jr., "The numerical solution of parabolic and elliptic differential equations," *Journal of the Society for Industrial and Applied Mathematics*, vol. 3, no. 1, pp. 28–41, 1955.

[20] X. Qin, S. Y. Cho, and L. Wang, "Convergence of splitting algorithms for the sum of two accretive operators with applications," *Fixed Point Theory Appl*, vol. 166, p. 2014, 2014.

[21] Y. Shehu, "Iterative approximations for zeros of sum of accretive operators in Banach spaces," *Journal of Function Spaces*, vol. 2016, Article ID 5973468, 9 pages, 2016.

[22] P. Tseng, "A modified forward-backward splitting method for maximal monotone mappings," *SIAM Journal on Control and Optimization*, vol. 38, no. 2, pp. 431–446, 2000.

[23] R. E. Bruck and S. Reich, "Nonexpansive projections and resolvents of accretive operators in Banach spaces," *Houston Journal of Mathematics*, vol. 3, pp. 459–470, 1977.

[24] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends® in Optimization*, vol. 1, pp. 123–231, 2013.

[25] F. Wang and H. Cui, "On the contraction-proximal point algorithms with multi-parameters," *Journal of Global Optimization*, vol. 54, no. 3, pp. 485–491, 2012.

[26] N. Xiu, D. Wang, and L. Kong, "A note on the gradient projection method with exact stepsize rule," *Journal of Computational Mathematics*, vol. 25, pp. 221–230, 2007.

[27] H.-K. Xu, "Averaged mappings and the gradient-projection algorithm," *Journal of Optimization Theory and Applications*, vol. 150, no. 2, pp. 360–378, 2011.

[28] Y. Yao, S. M. Kang, W. Jigang, and P.-X. Yang, "A regularized gradient projection method for the minimization problem," *Journal of Applied Mathematics*, vol. 2012, Article ID 259813, 9 pages, 2012.

[29] P. L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[30] H. H. Bauschke and P. L. Combettes, "A weak-to-strong convergence principle for fejér-monotone methods in Hilbert spaces," *Mathematics of Operations Research*, vol. 26, no. 2, pp. 248–264, 2001.

[31] J. Douglas and H. H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Transactions of the American Mathematical Society*, vol. 82, no. 2, p. 421, 1956.

[32] F. Alvarez and H. Attouch, "An inertial proximal method for monotone operators via discretization of a nonlinear oscillator with damping," *Set-Valued Analysis*, vol. 9, no. 1/2, pp. 3–11, 2001.

[33] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, NY, USA, 1987.

[34] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[35] Y. Dang, J. Sun, J. Sun, and H. Xu, "Inertial accelerated algorithms for solving a split feasibility problem," *Journal of Industrial & Management Optimization*, vol. 13, no. 3, pp. 1383–1394, 2017.

[36] Q. L. Dong, H. B. Yuan, Y. J. Cho, and T. M. Rassias, "Modified inertial Mann algorithm and inertial CQ-algorithm for nonexpansive mappings," *Optimization Letters*, vol. 12, no. 1, pp. 87–102, 2018.

[37] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $O(1/k^2)$," *Doklady Akademii Nauk: Archive*, vol. 269, pp. 543–547, 1983.

[38] A. Moudafi and M. Oliny, "Convergence of a splitting inertial proximal method for monotone operators," *Journal of Computational and Applied Mathematics*, vol. 155, no. 2, pp. 447–454, 2003.

[39] H. Attouch and A. Cabot, "Convergence of a relaxed inertial proximal algorithm for maximally monotone operators," *Mathematical Programming*, vol. 184, pp. 243–287, 2020.

[40] H. Attouch and A. Cabot, "Convergence rate of a relaxed inertial proximal algorithm for convex minimization," *Optimization*, vol. 69, no. 6, p. 1281, 2019.

[41] H. Attouch and A. Cabot, "Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions," *Applied Mathematics & Optimization*, vol. 80, no. 3, pp. 547–598, 2019.

[42] F. Iutzeler and J. M. Hendrickx, "A generic online acceleration scheme for optimization algorithms via relaxation and inertia," *Optimization Methods and Software*, vol. 34, no. 2, pp. 383–405, 2019.

[43] M. M. Alves and R. T. Marcavillaca, "On inexact relative-error hybrid proximal extragradient, forward-backward and tseng's modified forward-backward methods with inertial effects," *Set-Valued and Variational Analysis*, vol. 28, no. 2, pp. 301–325, 2019.

[44] M. M. Alves, J. Eckstein, M. Geremia, and J. G. Melo, "Relative-error inertial-relaxed inexact versions of Douglas-Rachford and ADMM splitting algorithms," *Computational Optimization and Applications*, vol. 75, no. 2, pp. 389–422, 2020.

[45] J. Abubakar, P. Kumam, A. Hassan Ibrahim, and A. Padcharoen, "Relaxed inertial tseng's type method for solving the inclusion problem with application to image restoration," *Mathematics*, vol. 8, no. 5, p. 818, 2020.

[46] W. Takahashi, Y. Takeuchi, and R. Kubota, "Strong convergence theorems by hybrid methods for families of nonexpansive mappings in Hilbert spaces," *Journal of Mathematical Analysis and Applications*, vol. 341, no. 1, pp. 276–286, 2008.

[47] W. R. Mann, "Mean value methods in iteration," *Proceedings of the American Mathematical Society*, vol. 4, no. 3, p. 506, 1953.

[48] J. Yang and H. Liu, "Strong convergence result for solving monotone variational inequalities in Hilbert space," *Numerical Algorithms*, vol. 80, no. 3, pp. 741–752, 2019.

[49] W. Takahashi, *Nonlinear Functional Analysis*, Yokohama Publishers, Yokohama, Japan, 2000.

[50] H. A. Hammad, W. Cholamjiak, and H. Dutta, "A modified shrinking projection methods for numerical reckoning fixed points of G-nonexpansive mappings in Hilbert spaces with graphs," *Miskolc Mathematical Notes*, vol. 20, no. 2, pp. 941–956, 2019.

[51] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, Springer, New York, NY, USA, 2011.

[52] C. Martinez-Yanes and H.-K. Xu, "Strong convergence of the CQ method for fixed point iteration processes," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 64, no. 11, pp. 2400–2411, 2006.

[53] D. A. Lorenz and T. Pock, "An inertial forward-backward algorithm for monotone inclusions," *Journal of Mathematical Imaging and Vision*, vol. 51, no. 2, pp. 311–325, 2015.

[54] Y. Censor and T. Elfving, "A multiprojection algorithm using Bregman projections in a product space," *Numerical Algorithms*, vol. 8, no. 2, pp. 221–239, 1994.

[55] Y. Censor, T. Elfving, N. Kopf, and T. Bortfeld, "The multiple-sets split feasibility problem and its applications for inverse problems," *Inverse Problems*, vol. 21, no. 6, pp. 2071–2084, 2005.

[56] C. Byrne, "Iterative oblique projection onto convex sets and the split feasibility problem," *Inverse Problems*, vol. 18, no. 2, pp. 441–453, 2002.

[57] C. Byrne, "A unified treatment of some iterative algorithms in signal processing and image reconstruction," *Inverse Problems*, vol. 20, no. 1, pp. 103–120, 2003.

[58] A. Gibali and D. V. Thong, "Tseng type methods for solving inclusion problems and its applications," *Calcolo*, vol. 55, no. 4, 2018.

[59] Y. Censor, T. Bortfeld, B. Martin, and A. Trofimov, "A unified approach for inversion problems in intensity-modulated

radiation therapy," *Physics in Medicine and Biology*, vol. 51, no. 10, pp. 2353–2365, 2006.

[60] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics. Springer, Berlin, Germany, 2011.

[61] Q. Dong, D. Jiang, P. Cholamjiak, and Y. Shehu, "A strong convergence result involving an inertial forward-backward algorithm for monotone inclusions," *Journal of Fixed Point Theory and Applications*, vol. 19, no. 4, pp. 3097–3118, 2017.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9767282, 14 pages
https://doi.org/10.1155/2020/9767282

Research Article

An Improved Differential Evolution Algorithm Based on Dual-Strategy

Xuxu Zhong [1] **and Peng Cheng** [2]

[1] *National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China*
[2] *School of Aeronautics and Astronautics, Sichuan University, Chengdu 610065, China*

Correspondence should be addressed to Peng Cheng; pengchengscu@163.com

Received 4 August 2020; Revised 23 September 2020; Accepted 9 October 2020; Published 2 November 2020

Academic Editor: Guoqiang Wang

In recent years, Differential Evolution (DE) has shown excellent performance in solving optimization problems over continuous space and has been widely used in many fields of science and engineering. How to avoid the local optimal solution and how to improve the convergence performance of DE are hotpot problems for many researchers. In this paper, an improved differential evolution algorithm based on dual-strategy (DSIDE) is proposed. The DSIDE algorithm has two strategies. (1) An enhanced mutation strategy based on "DE/rand/1," which takes into account the influence of reference individuals on mutation and has strong global exploration and convergence ability. (2) A novel adaptive strategy for scaling factor and crossover probability based on fitness value has a positive impact on population diversity. The DSIDE algorithm is verified with other seven state-of-the-art DE variants under 30 benchmark functions. Furthermore, Wilcoxon sign rank-sum test, Friedman test, and Kruskal–Wallis test are utilized to analyze the results. The experiment results show that the proposed DSIDE algorithm can significantly improve the global optimization performance.

1. Introduction

Differential Evolution (DE) is an emerging optimization technique proposed by Storn and Price [1] in 1995, which was initially used to solve Chebyshev polynomials. Later, it is demonstrated that DE is also an effective method to solve complex optimization problems. Similar to other intelligent evolutionary algorithms, DE is a stochastic parallel optimization algorithm based on swarm intelligence, which guides optimization search by imitating heuristic swarm intelligence generated by cooperation and competition among individuals in the population.

In DE, the population consists of several individuals, each of which representing a potential solution to an optimization problem. DE generates offspring individuals through mutation, crossover, and selection, and the offspring individuals are expected to be closer to the optimal solution. In the process of evolution, with the increase of generations, the population diversity becomes worse, leading to premature convergence or evolutionary stagnation, which is undoubtedly fatal to the algorithm that depends on the difference of population. Also, the performance of DE is affected by control parameters [2, 3]. For different optimization problems, these control parameters often need a large number of repeated experiments to adjust to the appropriate value for achieving better optimization effect.

To address these shortcomings in DE, many improvements have been proposed, most of which focused on control parameters and mutation strategies.

Population size NP, scaling factor F, and crossover probability CR are three crucial control parameters in DE. Experiments in many works of literatures show that the performance of DE can be improved by adjusting these control parameters. Omran et al. [4] proposed a self-adaptation scheme (SDE), in which F was adaptive and CR was generated by a normal distribution. Liu and Lampinen [5] proposed a fuzzy adaptive differential evolution algorithm (FADE), which used the fuzzy logic controller to adjust F, and CR

dynamically and successfully evolved individuals and their fitness values as input parameters of the logic controller. Brest et al. [6] developed a new adaptive DE algorithm, named jDE, applying F and CR to the individual level. If a better individual is produced, these parameters would be retained; otherwise, they would be adjusted according to two constants. Noman et al. [7] proposed an adaptive differential evolution algorithm (aDE), which was similar to jDE [6], except that the updating of parameters in aDE depended on whether the offspring was better than the average individual in the parent population. Asafuddoula et al. [8] used roulette to select the suitable CR value for each individual in each generation of the population. Tanabe and Fukunaga [9] proposed the success-history-based parameter adaptation for differential evolution (SHADE), which generated new F and CR pairs by sampling the nearby space of stored parameter pairs. Later, they came up with an improved version called L-SHADE [10]. Based on SHADE, a linear population size reduction strategy (LPSR) was adopted to reduce the population size NP by a linear function continuously. Zhu et al. [11] proposed an adaptive population tuning scheme (APTS) that dynamically adjusted the population size, in which redundant individuals were removed from the population or "excellent" individuals were generated. Zhao et al. [12] proposed a self-adaptive DE with population adjustment scheme (SAPA) to tune the size of the offspring population, which contained two kinds of population adjustment schemes. Pan et al. [13] proposed a parameter adaptive DE algorithm on real-parameter optimization, in which better control parameters F and CR are more likely to survive and produce good offspring. An enhancing DE with novel parameter control, referred to as DE-NPC, was proposed by Meng et al. [14]. The update of F and CR was based on the location information of the population and the success probability of CR, respectively, and a combined parabolic-linear population size reduction scheme was adopted. Di Carlo et al. [15] proposed a multipopulation adaptive version of inflationary DE algorithm (MP-AIDEA), the parameters F and CR of which were adjusted together with the local restart bubble size and the number of local restarts of Monotonic Basin Hopping [16]. Li et al. [17] presented an enhanced adaptive differential evolution algorithm (EJADE), in which CR sorting mechanism and dynamic population reduction strategy were introduced.

To improve the optimization performance and balance the contradiction between global exploration and local exploitation, researchers have carried out a lot of work on mutation strategy in DE. Das et al. [18] proposed an improved algorithm based on "DE/current-to-best/1" strategy, which made full use of the optimal individual information in the neighborhood to guide the mutation operation. Zhang and Sanderson [19] proposed an adaptive differential evolution algorithm (JADE), which adopted "DE/current-to-pbest/1" mutation model, used suboptimal solutions to improve population diversity, and employed Cauchy and Normal distribution to generate F and CR. Qin et al. [20] proposed a self-adaptive DE (SaDE), which adopted four mutation strategies to generate mutation individuals. The selection of mutation strategy would be affected by previous performance. A DE algorithm (CoDE) using three mutation strategies and three parameters for the random combination was presented by Wang et al. [21]. Epitropakis et al. [22] proposed a novel framework that specified the selection probability in the mutation operation based on the distance between each individual and the mutation individual, thereby guiding the population to global optimization. Mallipeddi et al. [23] proposed the EPSDE algorithm, which was characterized by a stochastic selection of mutation strategies and parameters in a candidate pool consisting of three basic mutation strategies and preset parameters. Xiang et al. [24] proposed an enhanced differential evolution algorithm (EDE), which adopted a new combined mutation strategy composed of "DE/current/1" and "DE/pbest/1." Cui et al. [25] proposed a DE algorithm based on adaptive multiple subgroups (MPADE), which divided the population into three subgroups according to fitness values, each subgroup had its mutation strategy. Wu et al. [26] presented a DE with multipopulation-based ensemble of mutation strategies (MPEDE), which had three mutation strategies, three indicator subgroups, and one reward subgroup. After several evolutionary generations, the reward subgroup was dynamically assigned to the best-performing mutation strategy. Parameters with an adaptive learning mechanism for the enhancement of differential evolution (PALM-DE) were presented by Meng et al. [27]. Unlike the external archive of the mutation strategy in JADE [19] and SHADE [9], the inferior solution archive in PALM-DE mutation strategy used a timestamp mechanism. In [28], Meng et al. introduced a novel parabolic population size reduction scheme and an enhanced timestamp-based mutation strategy to tackle the weakness of previous mutation strategy. Wei et al. [29] proposed the RPMDE algorithm, designed the "DE/M_pbest-best/1" mutation strategy, used the optimal individual group information to generate new solutions, and adopted the random perturbation method to avoid falling into the local optimal. Duan's DPLDE [30] algorithm used population diversity and population fitness to determine individuals participating in mutation operation, thus influencing the mutation strategy. Tian and Gao [31] proposed NDE, which employed two mutation operators based on neighborhood-based and an individual-based selection probability to adjust the search performance of each individual appropriately. Wang et al. [32] proposed the DE algorithm based on particle swarm optimization (DEPSO), which utilized the improved "DE/rand/1" mutation strategy and PSO mutation strategy. Meng and Pan [33] presented hierarchical archive based on mutation strategy with depth information of evolution for the enhancement of differential evolution (HARD-DE), the depth information in which was the linkage of more than three different generations of populations and was included into the mutation strategy. A hybrid differential evolution algorithm based on "DE/target-to-ci_mbest/1" mutation operation of CIPDE [34] and "DE/target-to-pbest/1" mutation operation of JADE [19] was introduced by Pan et al. [35]. Meng et al. [36] proposed depth information-based DE with adaptive parameter control (Di-DE), the mutation strategy of which contained a depth information-based external archive.

As mentioned above, mutation strategies and control parameters affect the performance of DE, and "DE/rand/1"

is widely used due to its strong global exploration ability and good population diversity. Many researchers have refined the mutation strategy. In this paper, an enhanced mutation strategy based on "DE/rand/1" is proposed by introducing a reference factor. Besides, according to the maximum, minimum, average fitness value of population, and the fitness value of the individual, the scaling factor and crossover probability are changed to adjust the population diversity effectively.

The remainder of the paper is organized as follows. Section 2 describes the basic DE algorithm. Section 3 provides the details of the proposed DSIDE. In Section 4, the proposed DSIDE is compared and analyzed experimentally with seven advanced DE algorithms, and the effectiveness of the enhanced mutation strategy and the novel adaptive strategy for control parameters in DSIDE is studied. Section 5 summarizes the work of this paper and points out the future research direction.

2. The Basic Differential Evolution Algorithm

An unconstrained optimization problem is to find the extremum of a function, which can be expressed as follows:

$$\begin{cases} \min & f(x_1, x_2, \ldots, x_D) \\ \text{s.t.} & x_j^L \le x_j \le x_j^U,\ j = 1, 2, \ldots, D, \end{cases} \tag{1}$$

where $f(*)$ denotes the fitness value, D represents the dimension of the problem, and x_j^L and x_j^U are the minimum and maximum values of x_j, respectively. The process of solving optimization problems in DE is divided into initialization, mutation, crossover, and selection.

2.1. Initialization. To establish a starting point, an initial population must be created in the search space. Without loss of generality, the jth component $(j = 1, 2, \ldots, D)$ of the ith individuals $(i = 1, 2, \ldots, \mathrm{NP})$ in the original population can be expressed as follows:

$$x_{i,j}^0 = x_{i,j}^L + \mathrm{rand}^*\left(x_{i,j}^U - x_{i,j}^L\right), \tag{2}$$

where rand returns a uniformly distributed random number between 0 and 1 and L and U represent the lower and upper bounds of solution space, respectively.

2.2. Mutation. The mutation strategy of the DE algorithm can be expressed as "DE/x/y," where "DE" means differential evolution algorithm, "x" represents the reference vector in the mutation operation, and "y" denotes the number of differential vectors in the mutation operation. The most common mutation strategy is to randomly select two different individuals in the population, scale their vector differences, and then conduct vector synthesis with another random individual. The obtained mutation individual V_i is as follows:

$$V_i^{G+1} = X_{r1}^G + F \cdot \left(X_{r2}^G - X_{r3}^G\right), \tag{3}$$

where $r1, r2$, and $r3$ are randomly generated integers ranging from 1 to NP, and $r1 \ne r2 \ne r3 \ne i$; G represents the current generation number; and F denotes the scaling factor and controls the amplification of the differential vector. The mutation strategy is shown in equation (3) and is known as "DE/rand/1".

2.3. Crossover. The purpose of the crossover operation is to generate the trial vector $U_{i,j}^{G+1}$. The binomial crossover and exponential crossover are two main crossover operators. In this paper, binomial crossover is adopted, and its expression is as follows:

$$U_{i,j}^{G+1} = \begin{cases} V_{i,j}^{G+1}, & \mathrm{rand} < \mathrm{CR} \text{ or } j = j_{\mathrm{rand}}, \\ X_{i,j}^G, & \text{otherwise}, \end{cases} \tag{4}$$

where $X_{i,j}^G$ denotes the jth component of the ith individual in the current population; $\mathrm{CR}(\in [0, 1])$ is called crossover probability, which determines the contribution of mutation vector $V_{i,j}^{G+1}$ to trial vector $U_{i,j}^{G+1}$. $j_{\mathrm{rand}}(\in [1, D])$ is a uniformly distributed random integer, ensuring that at least one-dimensional components of the trial vector $U_{i,j}^{G+1}$ inherit from the mutation vector $V_{i,j}^{G+1}$.

2.4. Selection. In DE, the greedy selection strategy is utilized to compare the trial vector $U_{i,j}^{G+1}$ with the target vector X_i^G, and the one which has better fitness value will be selected as the offspring individual X_i^{G+1}:

$$X_i^{G+1} = \begin{cases} U_i^{G+1}, & f\left(U_i^{G+1}\right) < f\left(X_i^G\right), \\ X_i^G, & \text{otherwise}, \end{cases} \tag{5}$$

where $f(\cdot)$ stands for the fitness value.

3. DSIDE Algorithm

In DSIDE, the crossover and selection operations are the same as the basic DE, as shown in equations (4) and (5), respectively. Next, the improved mutation strategy and adaptive strategy will be introduced.

3.1. An Enhanced Mutation Strategy. From equation (3), it can be seen that the reference individual X_{r1}^G plays an important role in regulating balance in the evolutionary process. In the early stage of evolution, when most individuals are far away from the optimal solution, a larger X_{r1}^G is conducive to jumping out of the local optimal. However, in the later stage of evolution, most individuals gradually approach the global optimal solution, and a larger X_{r1}^G may cause individuals to deviate from the correct direction of evolution, which is not in favor of global convergence. On this basis, we propose an improved mutation strategy as follows:

$$V_i^{G+1} = \alpha_i^{G*} X_{r1}^G + F_i \cdot \left(X_{r2}^G - X_{r3}^G\right), \tag{6}$$

$$\alpha_i^G = 1 - r^{\left(1 - G/G_{\max}\right)^2}. \tag{7}$$

In equation (6), α_i ($\in [0, 1]$), F_i, and CR_i are the reference factor, scaling factor, and crossover probability for each target individual X_i^G, respectively; G denotes the current generation number. In equation (7), r means a random number on the interval $[0, 1]$. $G_{\max}$ represents the maximum generation number. From equation (7) , it is not challenging to observe that the value of α_i^G is relatively large at the initial evolutionary stage, which ensures a wide range of search. As the evolutionary generation increases, the α_i^G value decreases and the search scope shrink.

3.2. A Novel Adaptive Strategy for Control Parameters. During the mutation operation of equation (3), the scaling factor affects the reference individual through the differential vector $(X_{r2}^G - X_{r3}^G)$, which is called "perturbation." A larger F can produce a larger "perturbation," which is helpful to maintain the population diversity, but will reduce the search efficiency of the algorithm. A smaller F helps to improve the convergence speed, but the loss of population diversity is faster, and it is easy to fall into local optimal and premature convergence. During the crossover operation of equation (4), CR determines the contribution of the mutation vector to trial vector. A larger CR facilitates the expansion of the search space, thus accelerating the convergence. However, the mutation individuals tend to be identical in the later evolutionary stage, which weights against the maintenance of diversity. A smaller CR is not to the benefit of exploring the search area. Therefore, F and CR should be adjusted adaptively to explore the global space more thoroughly in the early stage of evolution and exploit the local area near the optimal solution at the later stage of evolution. Based on these points, a novel adaptive strategy is proposed, which can dynamically adjust control parameters according to the fitness value, as shown in

$$F_i^G = \frac{\left(f_{\max}^G - f_i^G\right)}{f_{\mathrm{mean}}^G}, \tag{8}$$

$$\mathrm{CR}_i^G = \frac{\left(f_i^G - f_{\min}^G\right)}{f_{\mathrm{mean}}^G}, \tag{9}$$

where f_i^G is the fitness value of the target individual X_i^G, $f_{\max}^G$ and $f_{\min}^G$ are the maximum and minimum fitness values at the current generation G, and f_{mean}^G is the average fitness value of the current population.

The reference factor α_i^G, scaling factor F_i^G, and crossover probability CR_i^G are updated before each evolution. The entire process of DSIDE algorithm is shown in Algorithm 1.

```
Initialize the original population pop and calculate
  their fitness values, NP = 100, G = 1, Gmax=1000;
while ((G ≤ Gmax) do
  for each individual Xi in pop do
    Calculate αi in equation (7);
    Calculate Fi in equation (8);
    Calculate CRi in equation (9);
    Implement mutation in equation (6);
    Implement crossover in equation (4);
    Implement selection in equation (5);
  end for
  G = G + 1
end while
```

Algorithm 1: DSIDE.

4. Experimental Results and Analysis

4.1. Benchmark Functions. Unlike deterministic algorithms, it is difficult to verify that evolutionary algorithms are superior to other algorithms due to their limited knowledge. Therefore, benchmark functions are utilized to evaluate the performance of evolutionary algorithms. In this section, the performance of DSIDE is tested on 27 benchmark functions [37–39] listed in Table 1, where D is the dimension of the problem. $f_1 \sim f_{11}$ are unimodal functions. f_{12} has one minimum and is discontinuous. f_{13} is a noisy quadratic function. $f_{14} \sim f_{27}$ are multimodal functions. $f(*)$ denotes the global minimum value.

Experiment results in this paper are obtained on Windows 10 x64 Operating System of a PC with Intel (R) Core (TM) i7-8550U CPU (1.80 GHz) and 8 GB RAM, and algorithms are implemented in MATLAB 2015b Windows version.

4.2. Comparison with 7 Improved DE Algorithms. Here, we mainly discuss the overall optimization performance among jDE [6], JADE [19], SaDE [20], CoDE [21], EPSDE [23], MPEDE [26], DEPSO [32], and the proposed DSIDE algorithm. Experiments are carried out on $f_1 \sim f_{30}$ benchmark functions at 30 D and 100 D, respectively. The parameters of other algorithms are the same as in their original literatures. The population size NP is set to 100 for all algorithms. 30 independent runs with 1000 maximum number of evolutionary generations are conducted. Tables 2 and 3 show the mean/std (mean value and standard deviation) of fitness error over 30 runs at 30 D and 100 D, respectively. Symbols "+," "≈," and "−" behind "mean ± std" pair denote "Better Performance," "Similar Performance," and "Worse Performance," respectively, all of which are measured under Wilcoxon's signed-rank test with a level of significant $\alpha = 0.05$. Furthermore, Wilcoxon's rank-sum test and Kruskal–Wallis test [39, 40] in Tables 4–6 are employed to further test the optimization performance of all algorithms. The best results in tables are shown in bold. In addition, the representative convergence curves of all algorithms are also given in Figures 1 and 2.

Table 1: Benchmark functions.

Name	Function	Range	$f(*)$
Sphere	$f_1(x) = \sum_{i=1}^{D} x_i^2$	$[-100, 100]^D$	0
Elliptic	$f_2(x) = \sum_{i=1}^{D} (10^6)^{i-1/D-1} x_i^2$	$[-100, 100]^D$	0
Bent cigar	$f_3(x) = x_1^2 + 10^6 \sum_{i=2}^{D} x_i^2$	$[-100, 100]^D$	0
Schwefel 1.2	$f_4(x) = \sum_{i=1}^{D} (\sum_{j=1}^{i} x_j)^2$	$[-100, 100]^D$	0
Schwefel 2.22	$f_5(x) = \sum_{i=1}^{D} \lvert x_i \rvert + \prod_{i=1}^{D} \lvert x_i \rvert$	$[-10, 10]^D$	0
Schwefel 2.21	$f_6(x) = \max\{\lvert x_i \rvert, 1 \le i \le D\}$	$[-100, 100]^D$	0
Powell sum	$f_7(x) = \sum_{i=1}^{D} \lvert x_i \rvert^{i+1}$	$[-100, 100]^D$	0
Sum squares	$f_8(x) = \sum_{i=1}^{D} i x_i^2$	$[-10, 10]^D$	0
Discuss	$f_9(x) = 10^6 x_1^2 + \sum_{i=2}^{D} x_i^2$	$[-100, 100]^D$	0
Different powers	$f_{10}(x) = \sqrt{\sum_{i=1}^{D} \lvert x_i \rvert^{2+4(i-1/D-1)}}$	$[-100, 100]^D$	0
Zakharov	$f_{11}(x) = \sum_{i=1}^{D} x_i^2 + (\sum_{i=1}^{D} 0.5x_i)^2 + (\sum_{i=1}^{D} 0.5x_i)^4$	$[-5, 10]^D$	0
Step	$f_{12}(x) = \sum_{i=1}^{D} (\lvert x_i + 0.5 \rvert)^2$	$[-100, 100]^D$	0
Noise quartic	$f_{13}(x) = \sum_{i=1}^{D} i x_i^4 + \mathrm{rand}[0, 1)$	$[-1.28, 1.28]^D$	0
Rosenbrock	$f_{14}(x) = \sum_{i=1}^{D-1} [100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2]$	$[-30, 30]^D$	0
Griewank	$f_{15}(x) = \sum_{i=1}^{D} x_i^2/4000 - \prod_{i=1}^{D} \cos(x_i/\sqrt{i}) + 1$	$[-600, 600]^D$	0
Rastrigin	$f_{16}(x) = \sum_{i=1}^{D} (x_i^2 - 10\cos(2\pi x_i) + 10)$	$[-5.12, 5.12]^D$	0
Apline	$f_{17}(x) = \sum_{i=1}^{D} \lvert x_i \sin x_i + 0.1x_i \rvert$	$[-100, 100]^D$	0
Bohachevsky_2	$f_{18}(x) = \sum_{i=1}^{D-1} [x_i^2 + 2x_{i+1}^2 - 0.3\cos(3\pi x_i)\cos(3\pi x_{i+1}) + 0.3]$	$[-100, 100]^D$	0
Salomon	$f_{19}(x) = 1 - \cos(2\pi\sqrt{\sum_{i=1}^{D} x_i^2}) + 0.1\sqrt{\sum_{i=1}^{D} x_i^2}$	$[-100, 100]^D$	0
Scaffer2	$f_{20}(x) = \sum_{i=1}^{D} (x_i^2 + x_{i+1}^2)^{0.25} (\sin(50(x_i^2 + x_{i+1}^2)^{0.1}) + 1), x_{D+1} = x_1$	$[-100, 100]^D$	0
Weierstrass	$f_{21}(x) = \sum_{i=1}^{D} (\sum_{k=0}^{k_{\max}} [a^k \cos(2\pi b^k (x_i + 0.5))]) - D \sum_{k=0}^{k_{\max}} [a^k \cos(2\pi b^k \cdot 0.5)],\ a = 0.5,\ b = 3,\ k_{\max} = 20$	$[-0.5, 0.5]^D$	0
Katsuura	$f_{22}(x) = 10/D^2 \prod_{i=1}^{D} (1 + i\sum_{j=1}^{32} \lvert 2^j x_i - \mathrm{round}(2^j x_i) \rvert / 2^j)^{10/D^{1.2}} - 10/D^2$	$[-100, 100]^D$	0
HappyCat	$f_{23}(x) = \lvert \sum_{i=1}^{D} x_i^2 - D \rvert^{1/4} + (0.5\sum_{i=1}^{D} x_i^2 + \sum_{i=1}^{D} x_i)/D + 0.5$	$[-100, 100]^D$	0
HGBat	$f_{24}(x) = \lvert (\sum_{i=1}^{D} x_i^2)^2 - (\sum_{i=1}^{D} x_i)^2 \rvert^{1/2} + (0.5\sum_{i=1}^{D} x_i^2 + \sum_{i=1}^{D} x_i)/D + 0.5$	$[-100, 100]^D$	0
Scaffer's F6	$f_{25}(x) = \sum_{i=1}^{D} ((0.5 + ((\sqrt{x_i^2 + x_{i+1}^2}))^2 - 0.5)/(1 + 0.001(x_i^2 + x_{i+1}^2))^2),\ x_{D+1} = x_1$	$[-0.5, 0.5]^D$	0
Expanded Scaffer's F6	$f_{26}(x) = f_{25}(x_1, x_2) + f_{25}(x_2, x_3) + \cdots + f_{25}(x_{D-1}, x_D) + f_{25}(x_D, x_1)$	$[-5, 5]^D$	0
Expanded Griewank's plus Rosenbrock's	$f_{27}(x) = f_{15}(f_{14}(x_1, x_2)) + f_{15}(f_{14}(x_2, x_3)) + \cdots + f_{15}(f_{14}(x_{D-1}, x_D)) + f_{15}(f_{14}(x_D, x_1))$	$[-5.12, 5.12]^D$	0
NCRastrigin	$f_{28}(x) = \sum_{i=1}^{D} [y_i^2 - 10\cos(2\pi y_i) + 10],\ y_i = \begin{cases} x_i, \lvert x_i \rvert < 0.5 \\ \mathrm{round}(2x_i)/2, \lvert x_i \rvert \ge 0.5 \end{cases}$	$[-10, 10]^D$	0
Levy and Montalvo 1	$f_{29}(x) = \pi/D\{10(\sin(\pi y_1))^2 + \sum_{i=1}^{D-1} (y_i - 1)^2 [1 + 10(\sin(\pi y_{i+1}))^2] + (y_D - 1)^2\} + \sum_{i=1}^{D} u(x_i, 10, 100, 4)$	$[-10, 10]^D$	0
$y = 1 + 1/4(x_i + 1),\ u(x_i, a, k, m) = \begin{cases} k(x_i - a)^m, x_i > a \\ 0, -a \le x_i \le a \\ k(-x_i - a)^m, x_i < -a \end{cases}$		$[-10, 10]^D$	
Levy and Montalvo 2	$f_{30}(x) = 0.1\{10(\sin(3\pi x_1))^2 + \sum_{i=1}^{D-1} (x_i - 1)^2 [1 + (\sin(3\pi x_{i+1}))^2] + (x_D - 1)^2 [1 + (\sin(2\pi x_D))^2]\} + \sum_{i=1}^{D} u(x_i, 5, 100, 4)$	$[-5, 5]^D$	0

Table 2: Mean and STD obtained by jDE, JADE, SaDE, CoDE, EPSDE, MPEDE, DEPSO, and DSIDE on benchmark functions at 30 D.

F	jDE mean ± std	JADE mean ± std	SaDE mean ± std	CoDE mean ± std	EPSDE mean ± std	MPEDE mean ± std	DEPSO mean ± std	DSIDE mean ± std
f_1	$8.73E-18 \pm 7.55E-18$	$4.55E-36 \pm 2.17E-35$	$5.15E-24 \pm 4.88E-24$	$5.96E-08 \pm 2.50E-08$	$1.43E-26 \pm 4.42E-26$	$1.46E-29 \pm 7.94E-29$	$2.97E-99 \pm 1.50E-98$	$0.00E+00 \pm 0.00E+00$
f_2	$2.11E-14 \pm 1.51E-14$	$2.79E-31 \pm 1.28E-30$	$1.03E-20 \pm 1.53E-20$	$4.19E-05 \pm 1.66E-05$	$2.33E-22 \pm 3.90E-22$	$3.77E-27 \pm 1.66E-26$	$4.62E-96 \pm 2.52E-95$	$0.00E+00 \pm 0.00E+00$
f_3	$6.52E-12 \pm 4.85E-12$	$1.03E-30 \pm 2.97E-30$	$3.29E-18 \pm 2.78E-18$	$3.93E-02 \pm 1.40E-02$	$3.66E-21 \pm 9.09E-21$	$1.70E-25 \pm 6.43E-25$	$2.67E-87 \pm 1.46E-86$	$0.00E+00 \pm 0.00E+00$
f_4	$4.13E+00 \pm 5.42E+00$	$1.26E-13 \pm 1.41E-13$	$5.82E-01 \pm 4.56E-01$	$6.99E-02 \pm 3.65E-02$	$7.70E-01 \pm 4.18E+00$	$1.35E-06 \pm 7.28E-06$	$6.29E-89 \pm 3.25E-88$	$0.00E+00 \pm 0.00E+00$
f_5	$3.42E-11 \pm 1.34E-11$	$1.10E-14 \pm 3.28E-14$	$7.08E-14 \pm 3.16E-14$	$5.00E-05 \pm 9.51E-06$	$8.60E-11 \pm 1.21E-10$	$7.62E-16 \pm 1.87E-15$	$9.66E-53 \pm 3.75E-52$	$0.00E+00 \pm 0.00E+00$
f_6	$8.43E-01 \pm 8.20E-01$	$8.93E-12 \pm 1.79E-11$	$4.56E-02 \pm 7.63E-02$	$3.20E-01 \pm 8.09E-02$	$1.43E-01 \pm 1.44E-01$	$1.56E-07 \pm 1.71E-07$	$5.67E-47 \pm 3.11E-46$	$0.00E+00 \pm 0.00E+00$
f_7	$3.10E-41 \pm 1.55E-40$	$9.36E-44 \pm 4.29E-43$	$4.48E-22 \pm 1.81E-21$	$2.14E-17 \pm 5.29E-17$	$3.95E-38 \pm 2.05E-37$	$6.15E-53 \pm 3.29E-52$	$1.59E-115 \pm 8.70E-115$	$0.00E+00 \pm 0.00E+00$
f_8	$1.37E-18 \pm 1.36E-18$	$2.92E-37 \pm 1.02E-36$	$6.13E-25 \pm 5.88E-25$	$6.68E-09 \pm 2.67E-09$	$3.34E-28 \pm 5.36E-28$	$4.89E-32 \pm 2.63E-31$	$4.93E-98 \pm 2.70E-97$	$0.00E+00 \pm 0.00E+00$
f_9	$1.78E-17 \pm 2.03E-17$	$1.43E-32 \pm 7.67E-32$	$1.25E-23 \pm 1.07E-23$	$8.20E-08 \pm 3.35E-08$	$1.21E-25 \pm 2.64E-25$	$9.10E-29 \pm 4.54E-28$	$2.06E-98 \pm 1.01E-97$	$0.00E+00 \pm 0.00E+00$
f_{10}	$9.12E-13 \pm 5.83E-13$	$1.07E-23 \pm 3.33E-23$	$1.88E-12 \pm 2.77E-12$	$5.14E-06 \pm 1.63E-06$	$1.35E-18 \pm 1.90E-18$	$1.13E-20 \pm 3.84E-20$	$5.52E-55 \pm 2.89E-54$	$0.00E+00 \pm 0.00E+00$
f_{11}	$2.25E-15 \pm 3.04E-15$	$1.27E-39 \pm 6.75E-39$	$1.17E-21 \pm 1.14E-21$	$2.06E-08 \pm 9.16E-09$	$1.40E-26 \pm 5.53E-26$	$7.92E-34 \pm 2.01E-33$	$5.09E-104 \pm 1.33E-103$	$0.00E+00 \pm 0.00E+00$
f_{12}	$9.58E-18 \pm 1.04E-17$	$3.08E-34 \pm 1.24E-33$	$4.94E-24 \pm 6.24E-24$	$6.13E-08 \pm 1.49E-08$	$5.99E-27 \pm 1.74E-26$	$2.12E-31 \pm 6.85E-31$	$2.73E+00 \pm 2.69E-01$	$1.18E+00 \pm 2.03E-01$
f_{13}	$1.18E-02 \pm 3.27E-03$	$3.98E-03 \pm 1.54E-03$	$4.87E-03 \pm 1.87E-03$	$1.38E-02 \pm 3.77E-03$	$4.77E-03 \pm 2.22E-03$	$3.44E-03 \pm 1.61E-03$	$3.43E-01 \pm 2.39E-01$	$8.49E-04 \pm 7.45E-04$
f_{14}	$2.66E+01 \pm 1.36E+01$	$1.65E-01 \pm 8.37E-01$	$2.90E+01 \pm 1.44E+01$	$2.23E+01 \pm 5.32E-01$	$1.04E+01 \pm 3.21E+00$	$3.13E+00 \pm 4.07E+00$	$2.80E+01 \pm 2.77E-01$	$2.85E+01 \pm 9.98E-02$
f_{15}	$0.00E+00 \pm 0.00E+00$	$1.78E-11 \pm 9.73E-11$	$1.64E-03 \pm 4.26E-03$	$1.82E-05 \pm 4.47E-05$	$0.00E+00 \pm 0.00E+00$	$1.40E-03 \pm 3.76E-03$	$7.66E-04 \pm 4.20E-03$	$0.00E+00 \pm 0.00E+00$
f_{16}	$1.47E-04 \pm 7.04E-04$	$1.87E-04 \pm 8.99E-05$	$5.32E-01 \pm 8.10E-01$	$2.46E+01 \pm 1.94E+00$	$6.29E-01 \pm 7.87E-01$	$2.78E-12 \pm 6.03E-12$	$3.94E+00 \pm 2.14E+01$	$0.00E+00 \pm 0.00E+00$
f_{17}	$1.47E-03 \pm 4.72E-04$	$1.13E-02 \pm 3.83E-03$	$8.25E-04 \pm 3.77E-04$	$3.09E+01 \pm 3.80E+00$	$1.42E-02 \pm 5.22E-03$	$2.27E-07 \pm 1.24E-06$	$1.57E-51 \pm 4.78E-51$	$0.00E+00 \pm 0.00E+00$
f_{18}	$1.30E-16 \pm 2.11E-16$	$0.00E+00 \pm 0.00E+00$	$9.34E-02 \pm 3.08E-01$	$4.32E-06 \pm 2.06E-06$	$0.00E+00 \pm 0.00E+00$	$3.98E-02 \pm 2.18E-01$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{19}	$2.15E-01 \pm 3.45E-02$	$2.03E-01 \pm 1.83E-02$	$2.13E-01 \pm 3.46E-02$	$3.64E-01 \pm 4.62E-02$	$1.68E-01 \pm 4.68E-02$	$2.57E-01 \pm 5.68E-02$	$9.99E-02 \pm 1.53E-07$	$0.00E+00 \pm 0.00E+00$
f_{20}	$5.68E+00 \pm 1.32E+00$	$4.63E+00 \pm 7.82E-01$	$1.64E+01 \pm 2.68E+00$	$2.05E+01 \pm 2.97E+00$	$1.19E+01 \pm 1.68E+00$	$2.26E+00 \pm 1.13E+00$	$3.83E-01 \pm 1.01E+00$	$0.00E+00 \pm 0.00E+00$
f_{21}	$3.38E-10 \pm 3.98E-10$	$2.40E-05 \pm 3.93E-05$	$7.41E-14 \pm 8.66E-14$	$1.26E-02 \pm 1.57E-03$	$6.42E+00 \pm 2.36E+00$	$4.38E-03 \pm 1.48E-02$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{22}	$5.66E-02 \pm 8.32E-03$	$2.65E-03 \pm 3.29E-04$	$1.45E-01 \pm 1.92E-02$	$8.84E-03 \pm 7.08E-04$	$8.85E-02 \pm 1.32E-02$	$6.99E-04 \pm 1.43E-04$	$6.16E-01 \pm 9.64E-02$	$0.00E+00 \pm 0.00E+00$
f_{23}	$3.66E-01 \pm 5.20E-02$	$2.53E-01 \pm 3.62E-02$	$3.59E-01 \pm 6.11E-02$	$4.85E-01 \pm 4.78E-02$	$3.30E-01 \pm 3.86E-02$	$2.76E-01 \pm 5.31E-02$	$8.49E-01 \pm 9.02E-02$	$5.04E-01 \pm 6.36E-02$
f_{24}	$3.39E-01 \pm 3.77E-02$	$3.71E-01 \pm 1.28E-01$	$4.05E-01 \pm 1.15E-01$	$2.96E-01 \pm 2.72E-02$	$3.41E-01 \pm 7.43E-02$	$4.11E-01 \pm 1.61E-01$	$4.98E-01 \pm 6.20E-03$	$4.88E-01 \pm 1.51E-02$
f_{25}	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$	$2.43E-12 \pm 1.10E-12$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{26}	$7.91E-01 \pm 1.37E-01$	$6.77E-01 \pm 6.13E-02$	$1.21E+00 \pm 1.46E-01$	$1.70E+00 \pm 1.77E-01$	$2.04E+00 \pm 2.45E-01$	$3.45E-01 \pm 7.08E-02$	$4.63E+00 \pm 8.45E-01$	$0.00E+00 \pm 0.00E+00$
f_{27}	$3.24E+00 \pm 3.45E-01$	$2.83E+00 \pm 2.07E-01$	$6.28E+00 \pm 5.62E-01$	$8.47E+00 \pm 6.77E-01$	$4.59E+00 \pm 4.34E-01$	$2.11E+00 \pm 2.55E-01$	$1.17E+01 \pm 4.46E-01$	$1.18E+01 \pm 4.67E-01$
f_{28}	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$	$8.00E+00 \pm 1.20E+00$	$4.84E-06 \pm 2.39E-05$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{29}	$3.91E-21 \pm 3.66E-21$	$1.57E-32 \pm 5.57E-48$	$2.01E-27 \pm 2.69E-27$	$6.64E-11 \pm 3.04E-11$	$2.57E-22 \pm 9.18E-22$	$1.57E-32 \pm 5.57E-48$	$8.21E-02 \pm 3.15E-02$	$2.78E-03 \pm 1.82E-03$
f_{30}	$2.25E-20 \pm 3.06E-20$	$5.83E-29 \pm 2.53E-28$	$6.30E-26 \pm 7.49E-26$	$1.52E-10 \pm 7.18E-11$	$1.05E-23 \pm 2.34E-23$	$9.40E-30 \pm 4.89E-29$	$3.29E-01 \pm 7.20E-02$	$6.19E-01 \pm 2.04E-01$
−	20	20	22	23	20	21	23	
≈	3	3	2	0	3	2	4	
+	7	7	6	7	7	7	3	

Table 3: Mean and STD obtained by jDE, JADE, SaDE, CoDE, EPSDE, MPEDE, DEPSO, and DSIDE on benchmark functions at 100 D.

F	jDE mean ± std	JADE mean ± std	SaDE mean ± std	CoDE mean ± std	EPSDE mean ± std	MPEDE mean ± std	DEPSO mean ± std	DSIDE mean ± std
f_1	$1.62E-05 \pm 7.35E-06$	$3.01E-16 \pm 3.28E-16$	$2.97E-04 \pm 1.36E-04$	$7.40E-01 \pm 2.67E-01$	$2.17E-08 \pm 2.20E-08$	$9.24E-10 \pm 6.63E-10$	$5.93E-95 \pm 2.86E-94$	$0.00E+00 \pm 0.00E+00$
f_2	$8.40E-02 \pm 5.79E-02$	$4.17E-11 \pm 4.24E-11$	$9.04E-01 \pm 5.30E-01$	$2.14E+03 \pm 6.19E+02$	$2.40E-04 \pm 2.59E-04$	$1.46E-04 \pm 1.74E-04$	$2.58E-84 \pm 1.41E-83$	$0.00E+00 \pm 0.00E+00$
f_3	$1.57E+01 \pm 7.77E+00$	$1.21E-09 \pm 1.60E-09$	$3.41E+02 \pm 2.25E+02$	$7.25E+05 \pm 1.84E+05$	$2.02E-02 \pm 1.25E-02$	$5.18E-03 \pm 1.21E-02$	$8.14E-84 \pm 4.46E-83$	$0.00E+00 \pm 0.00E+00$
f_4	$8.14E+03 \pm 2.80E+03$	$2.13E+02 \pm 1.02E+02$	$1.59E+03 \pm 3.08E+02$	$1.42E+03 \pm 3.22E+02$	$2.96E+04 \pm 3.07E+04$	$2.89E+02 \pm 1.33E+02$	$6.91E-89 \pm 3.76E-88$	$0.00E+00 \pm 0.00E+00$
f_5	$7.69E-04 \pm 3.31E-04$	$3.26E-08 \pm 3.18E-08$	$1.26E-03 \pm 1.75E-03$	$1.17E+00 \pm 2.04E-01$	$8.04E-05 \pm 1.17E-04$	$3.84E-03 \pm 1.92E-02$	$2.40E-49 \pm 1.07E-48$	$0.00E+00 \pm 0.00E+00$
f_6	$2.70E+01 \pm 4.36E+00$	$7.18E+00 \pm 9.65E-01$	$1.18E+01 \pm 1.82E+00$	$5.00E+00 \pm 7.39E-01$	$6.70E+01 \pm 1.87E+01$	$1.00E+01 \pm 1.56E+00$	$1.06E-45 \pm 3.44E-45$	$0.00E+00 \pm 0.00E+00$
f_7	$3.56E+27 \pm 1.80E+28$	$5.29E+33 \pm 2.67E+34$	$9.40E+53 \pm 5.13E+54$	$8.33E+20 \pm 3.17E+21$	$1.08E+41 \pm 5.36E+41$	$1.41E+38 \pm 7.57E+38$	$1.14E-114 \pm 6.23E-114$	$0.00E+00 \pm 0.00E+00$
f_8	$7.78E-06 \pm 3.71E-06$	$2.06E-16 \pm 2.78E-16$	$1.15E-04 \pm 5.85E-05$	$2.89E-01 \pm 6.72E-02$	$6.55E-09 \pm 3.47E-09$	$8.29E-10 \pm 6.98E-10$	$4.95E-97 \pm 2.64E-96$	$0.00E+00 \pm 0.00E+00$
f_9	$2.91E-05 \pm 1.29E-05$	$9.47E-16 \pm 7.48E-16$	$5.46E-04 \pm 2.94E-04$	$9.39E-01 \pm 2.85E-01$	$2.08E-07 \pm 1.08E-07$	$6.02E-09 \pm 7.61E-09$	$2.33E-99 \pm 6.98E-99$	$0.00E+00 \pm 0.00E+00$
f_{10}	$1.93E-02 \pm 3.38E-02$	$2.56E-08 \pm 1.84E-08$	$2.86E-01 \pm 9.05E-02$	$1.49E+00 \pm 4.09E-01$	$3.43E-03 \pm 2.80E-03$	$1.23E-05 \pm 7.38E-06$	$2.95E-54 \pm 1.16E-53$	$0.00E+00 \pm 0.00E+00$
f_{11}	$1.24E-03 \pm 7.08E-04$	$8.30E-16 \pm 6.38E-16$	$2.22E-04 \pm 1.27E-04$	$1.86E-02 \pm 5.83E-03$	$7.82E-06 \pm 8.22E-06$	$2.63E-09 \pm 2.86E-09$	$1.54E-95 \pm 8.42E-95$	$0.00E+00 \pm 0.00E+00$
f_{12}	$1.96E-05 \pm 9.68E-06$	$2.58E-16 \pm 2.30E-16$	$2.84E-04 \pm 1.37E-04$	$7.84E-01 \pm 2.39E-01$	$1.58E-08 \pm 8.52E-09$	$9.56E-10 \pm 1.02E-09$	$1.85E+01 \pm 5.49E-01$	$1.19E+01 \pm 5.90E-01$
f_{13}	$7.17E-02 \pm 1.28E-02$	$4.21E-02 \pm 9.86E-03$	$1.31E-01 \pm 2.88E-02$	$6.42E-02 \pm 1.71E-02$	$5.16E-02 \pm 2.20E-02$	$7.10E-02 \pm 1.96E-02$	$3.31E-01 \pm 1.86E-01$	$9.70E-04 \pm 1.13E-03$
f_{14}	$1.97E+02 \pm 5.74E+01$	$1.47E+02 \pm 4.80E+01$	$4.60E+02 \pm 8.04E+01$	$2.07E+02 \pm 3.88E+01$	$1.92E+02 \pm 5.14E+01$	$1.77E+02 \pm 6.47E+01$	$9.85E+01 \pm 2.12E-01$	$9.83E+01 \pm 7.31E-02$
f_{15}	$9.69E-06 \pm 4.04E-06$	$2.46E-03 \pm 7.91E-03$	$7.70E-03 \pm 1.24E-02$	$4.16E-01 \pm 1.29E-01$	$1.31E-03 \pm 3.59E-03$	$4.92E-03 \pm 9.66E-03$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{16}	$1.83E+02 \pm 1.53E+01$	$1.56E+02 \pm 9.95E+00$	$2.61E+02 \pm 1.33E+01$	$7.14E+02 \pm 2.16E+01$	$4.34E+02 \pm 2.70E+01$	$4.35E+01 \pm 8.85E+00$	$3.32E-02 \pm 1.82E-01$	$0.00E+00 \pm 0.00E+00$
f_{17}	$1.85E+00 \pm 2.21E+00$	$1.58E+01 \pm 5.40E+00$	$3.00E+00 \pm 4.15E+00$	$8.86E+01 \pm 1.22E+01$	$1.63E+01 \pm 2.63E+01$	$2.49E+00 \pm 2.36E+00$	$4.52E-50 \pm 1.48E-49$	$0.00E+00 \pm 0.00E+00$
f_{18}	$6.15E-04 \pm 3.10E-04$	$2.35E+00 \pm 1.45E+00$	$7.00E+00 \pm 2.71E+00$	$2.06E+01 \pm 3.65E+00$	$2.45E+00 \pm 2.02E+00$	$5.08E+00 \pm 2.46E+00$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{19}	$9.83E-01 \pm 8.78E-02$	$7.00E-01 \pm 1.02E-01$	$1.63E+00 \pm 2.27E-01$	$1.90E+00 \pm 1.39E-01$	$1.08E+00 \pm 1.84E-01$	$1.39E+00 \pm 2.82E-01$	$9.99E-02 \pm 1.54E-07$	$0.00E+00 \pm 0.00E+00$
f_{20}	$9.07E+01 \pm 9.50E+00$	$8.89E+01 \pm 6.90E+00$	$1.57E+02 \pm 2.02E+01$	$3.01E+02 \pm 9.12E+00$	$2.95E+02 \pm 2.55E+01$	$1.25E+01 \pm 4.92E+00$	$4.57E-02 \pm 1.44E-01$	$0.00E+00 \pm 0.00E+00$
f_{21}	$5.75E-02 \pm 1.55E-02$	$2.79E+00 \pm 1.49E+00$	$2.83E+00 \pm 1.32E+00$	$6.06E+00 \pm 7.75E-01$	$1.06E+02 \pm 2.95E+00$	$1.35E+01 \pm 2.95E+00$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{22}	$2.20E-01 \pm 2.56E-02$	$1.08E-01 \pm 9.17E-03$	$5.49E-01 \pm 3.87E-02$	$4.01E-01 \pm 3.50E-02$	$5.82E-01 \pm 3.74E-02$	$9.35E-03 \pm 1.69E-03$	$2.11E+00 \pm 1.80E-01$	$0.00E+00 \pm 0.00E+00$
f_{23}	$6.32E-01 \pm 5.38E-02$	$5.17E-01 \pm 6.27E-02$	$6.27E-01 \pm 7.99E-02$	$7.20E-01 \pm 8.36E-02$	$5.81E-01 \pm 5.89E-02$	$5.76E-01 \pm 7.39E-02$	$1.19E+00 \pm 1.38E-01$	$8.72E-01 \pm 4.53E-02$
f_{24}	$5.05E-01 \pm 1.73E-01$	$5.03E-01 \pm 2.03E-01$	$6.10E-01 \pm 1.98E-01$	$6.39E-01 \pm 2.36E-01$	$5.74E-01 \pm 1.96E-01$	$5.46E-01 \pm 2.32E-01$	$5.00E-01 \pm 6.71E-11$	$4.99E-01 \pm 7.93E-04$
f_{25}	$1.07E-09 \pm 7.45E-10$	$1.48E-16 \pm 3.16E-16$	$1.56E-08 \pm 8.66E-09$	$3.55E-05 \pm 1.17E-05$	$9.54E-13 \pm 6.18E-13$	$1.02E-13 \pm 1.81E-13$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{26}	$1.00E+01 \pm 8.76E-01$	$9.01E+00 \pm 3.94E-01$	$1.55E+01 \pm 6.55E-01$	$2.28E+01 \pm 8.47E-01$	$2.75E+01 \pm 1.97E+00$	$2.36E+00 \pm 7.01E-01$	$2.71E+01 \pm 1.09E+01$	$0.00E+00 \pm 0.00E+00$
f_{27}	$3.33E+01 \pm 3.12E+00$	$2.88E+01 \pm 1.74E+00$	$4.99E+01 \pm 2.06E+00$	$7.00E+01 \pm 2.17E+00$	$6.41E+01 \pm 7.46E+00$	$1.84E+01 \pm 2.96E+00$	$4.41E+01 \pm 3.76E-01$	$4.43E+01 \pm 5.39E-01$
f_{28}	$5.61E+01 \pm 1.07E+01$	$5.03E+01 \pm 6.74E+00$	$1.03E+02 \pm 1.17E+01$	$6.24E+02 \pm 3.48E+01$	$2.60E+02 \pm 2.70E+01$	$3.32E+01 \pm 6.76E+00$	$0.00E+00 \pm 0.00E+00$	$0.00E+00 \pm 0.00E+00$
f_{29}	$6.30E-09 \pm 4.06E-09$	$1.95E-20 \pm 2.12E-20$	$8.31E-09 \pm 3.84E-09$	$6.26E-05 \pm 2.09E-05$	$7.85E-13 \pm 8.09E-13$	$2.07E-03 \pm 7.89E-03$	$3.76E-01 \pm 6.85E-02$	$2.23E-01 \pm 3.48E-02$
f_{30}	$3.54E-08 \pm 2.69E-08$	$3.16E-17 \pm 8.92E-17$	$7.03E-03 \pm 1.15E-02$	$2.24E-03 \pm 1.18E-03$	$3.33E-03 \pm 5.17E-03$	$9.99E-03 \pm 2.17E-02$	$4.14E+00 \pm 6.81E-01$	$7.38E+00 \pm 4.33E-01$
−	25	25	26	26	26	25	23	
≈	0	0	0	0	0	0	5	
+	5	5	4	4	4	5	2	

Table 4: The results of Wilcoxon's rank-sum test over independent 30 runs.

Comparison	R^+	$D = 30$ R^-	p value	$\alpha = 0.1$	R^+	R^-	p value	$D = 100$ $\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.1$
DSIDE vs. jDE	235	143	$6.37e-04$	Yes	Yes	372	93	$6.01e-06$	Yes	Yes
DSIDE vs. JADE	223	155	$1.53e-03$	Yes	Yes	371	94	$6.94e-06$	Yes	Yes
DSIDE vs. SaDE	282	124	$2.84e-04$	Yes	Yes	402	63	$1.37e-06$	Yes	Yes
DSIDE vs CoDE	322	143	$3.15e-05$	Yes	Yes	419	46	$1.76e-07$	Yes	Yes
DSIDE vs. EPSDE	240	138	$3.80e-04$	Yes	Yes	401	64	$2.00e-06$	Yes	Yes
DSIDE vs. MPEDE	244	162	$7.53e-04$	Yes	Yes	369	96	$6.46e-06$	Yes	Yes
DSIDE vs. DEPSO	294	57	$7.51e-04$	Yes	Yes	284	41	$1.69e-03$	Yes	Yes

Table 5: The results of Friedman and Kruskal–Wallis tests on 30**D** test functions.

Algorithms	jDE	JADE	SaDE	CoDE	EPSDE	MPEDE	DEPSO	DSIDE
Friedman (rank)	5.25	3.12	5.42	6.73	4.92	3.53	4.28	2.75
Kruskal–Wallis (rank)	131.13	109.28	133.87	163.00	129.93	117.38	112.73	66.67

Table 6: The results of Friedman and Kruskal–Wallis tests on 100 D test functions.

Algorithms	jDE	JADE	SaDE	CoDE	EPSDE	MPEDE	DEPSO	DSIDE
Friedman (rank)	4.80	3.10	6.23	6.77	5.50	4.23	3.32	2.05
Kruskal–Wallis (rank)	132.77	120.80	146.30	165.20	137.87	128.70	81.03	51.33

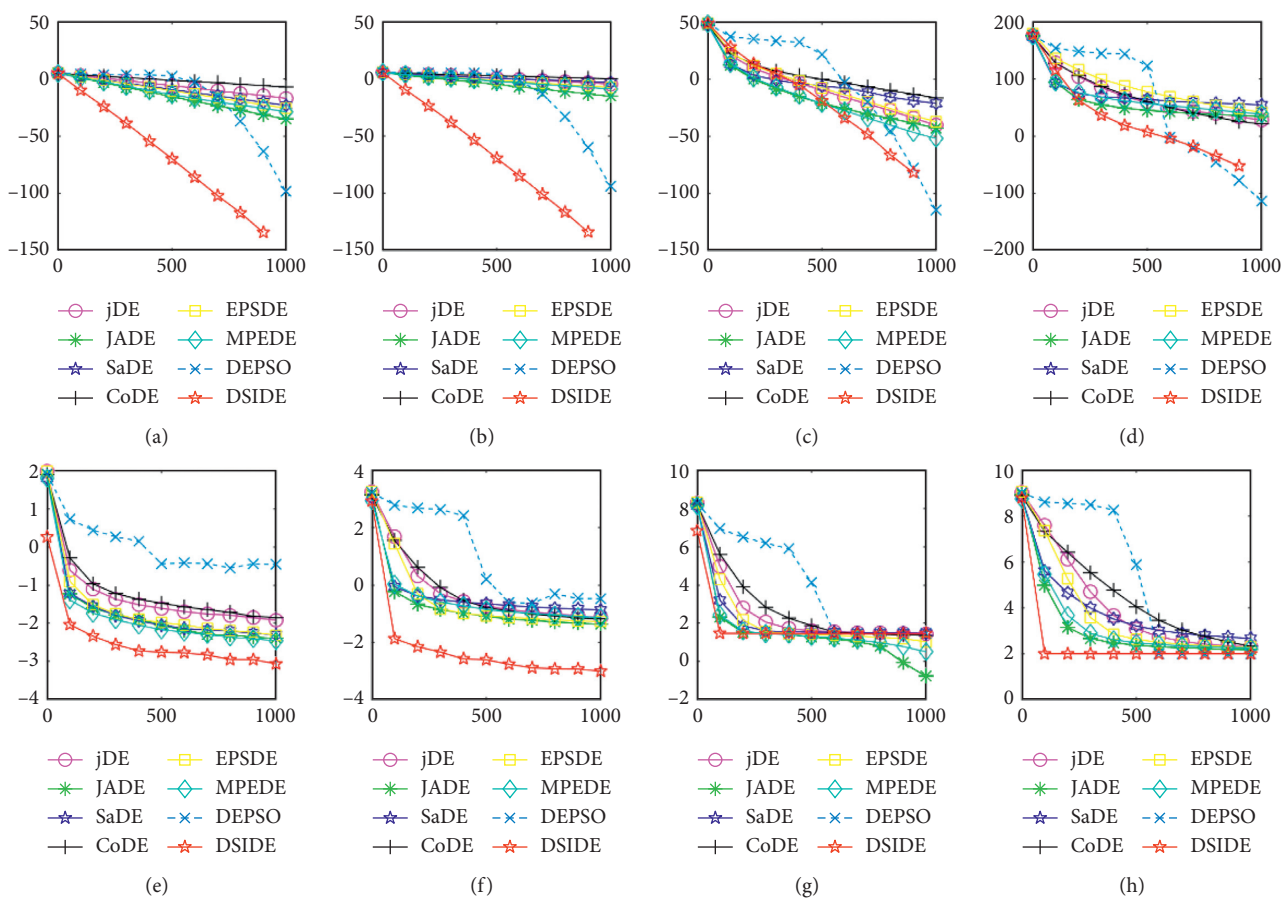


Figure 1: Continued.

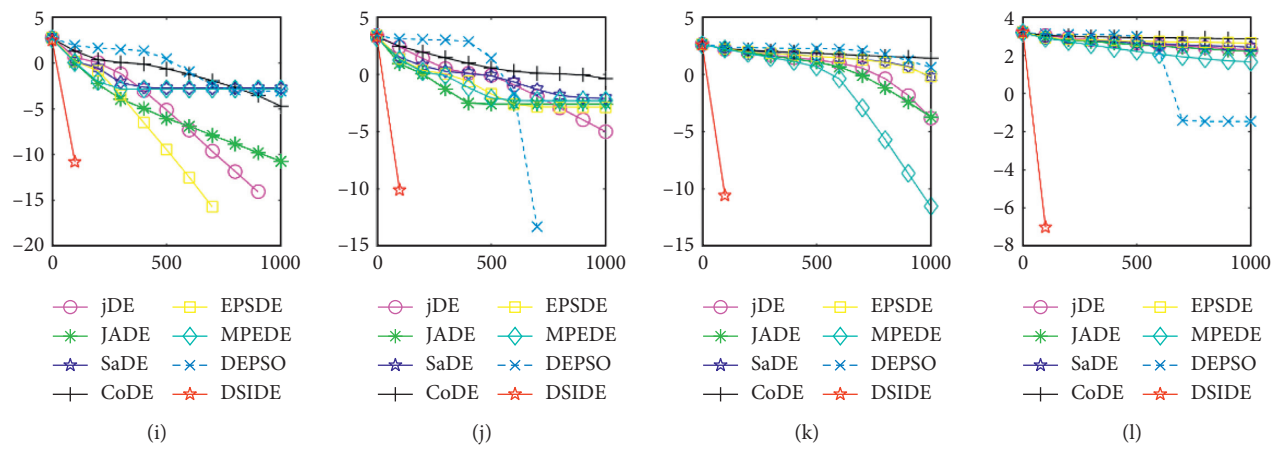


Figure 1: Convergence curves of $f_1, f_7, f_{13}, f_{14}, f_{15}$, and f_{16} at $D = 30,100$. The horizontal axis and the vertical axis are generations and the mean function error values over 30 independent runs. (a) f_1 30 D, (b) f_1 100 D, (c) f_7 30 D, (d) f_7 100 D, (e) f_{13} 30 D, (f) f_{13} 100 D, (g) f_{14} 30 D, (h) f_{14} 100 D, (i) f_{15} 30 D, (j) f_{15} 100 D, (k) f_{16} 30 D, and (l) f_{16} 100 D.

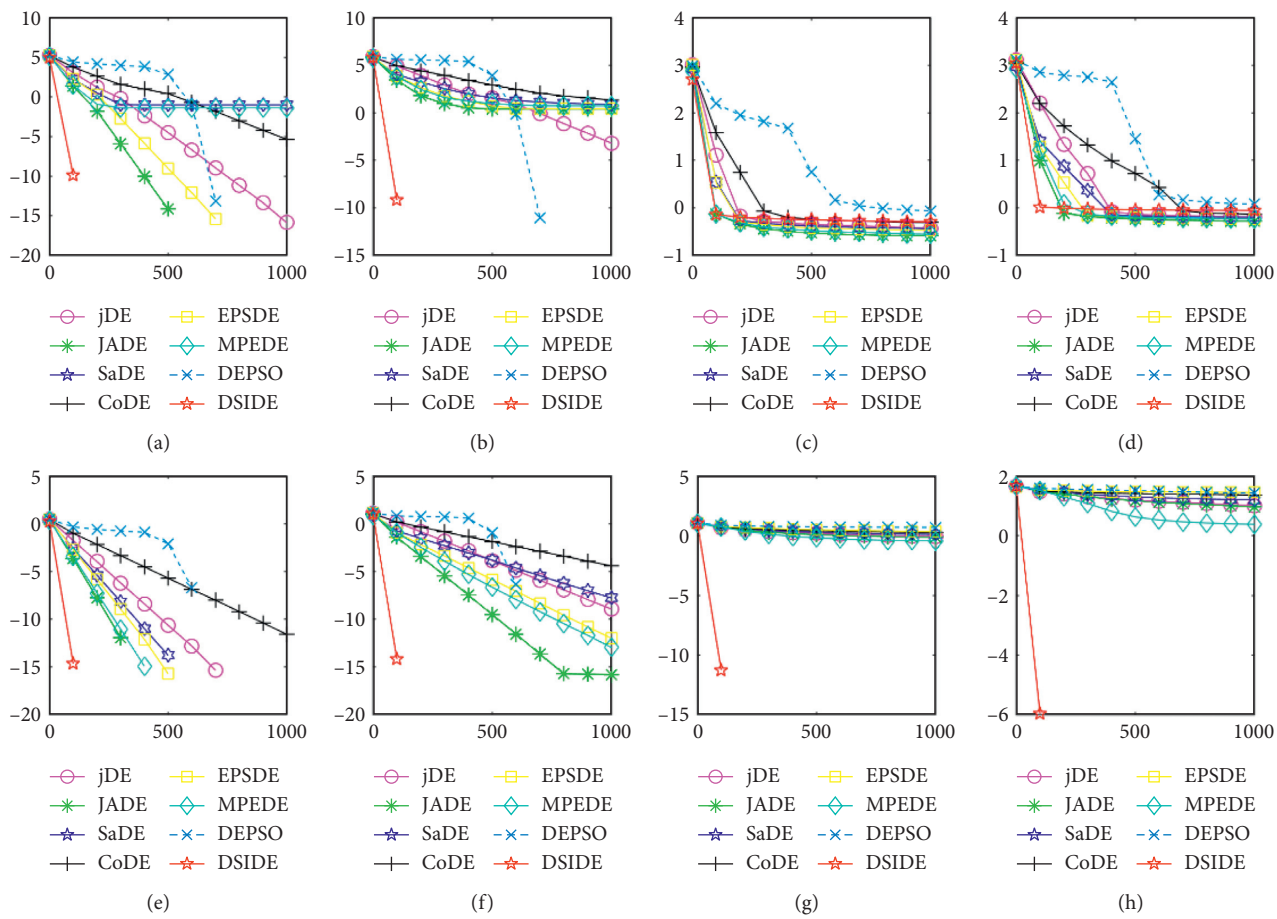


Figure 2: Continued.

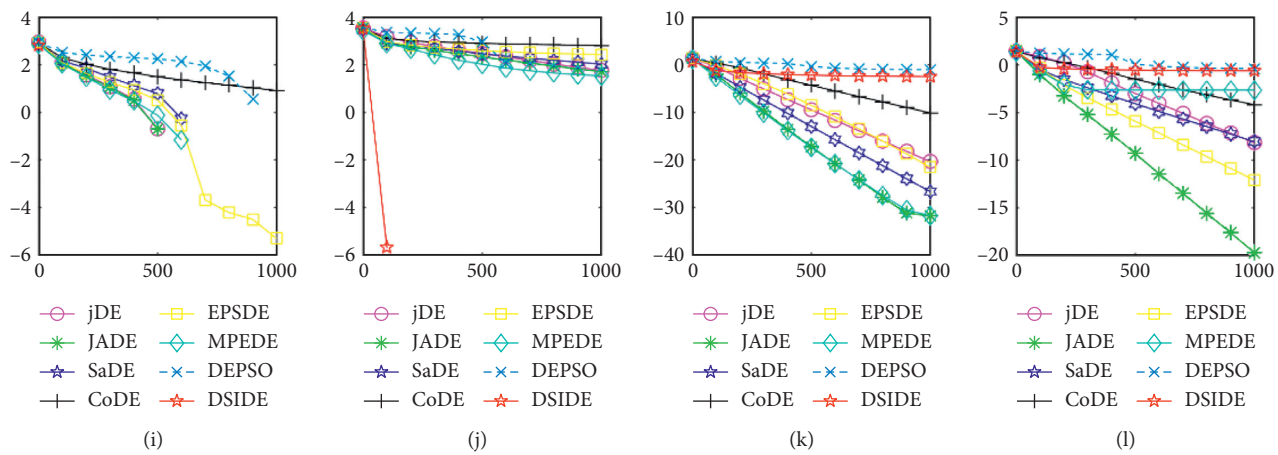


Figure 2: Convergence curves of f_{18}, f_{23}, f_{25}, f_{26}, f_{28}, and f_{29} at $D = 30,100$. The horizontal axis and the vertical axis are generations and the mean function error values over 30 independent runs. (a) f_{18} 30 D, (b) f_{18} 100 D, (c) f_{23} 30 D, (d) f_{23} 100 D, (e) f_{25} 30 D, (f) f_{25} 100 D, (g) f_{26} 30 D, (h) f_{26} 100 D, (i) f_{28} 30 D, (j) f_{28} 100 D, (k) f_{29} 30 D, and (l) f_{29} 100 D.

From Table 2 on 30 D, the proposed DSIDE algorithm displays that 23 out of 30 benchmark functions have better or similar performance than jDE, JADE, CoDE, EPSDE, and MPEDE, 24 out of 30 benchmark functions have better or comparable performance than SaDE, and 27 out of 30 benchmark functions have better or equivalent performance than DEPSO. Furthermore, the proposed DSIDE algorithm performs the best on benchmark functions $f_1 \sim f_{11}$, f_{13}, $f_{15} \sim f_{22}$, $f_{25} \sim f_{26}$, and f_{28} in comparison with the other contrasted algorithms, performs slightly inferior on benchmark functions f_{14}, f_{23}, f_{24}, and f_{27}, and only performs poorly on f_{12}, f_{29}, and f_{30}. Therefore, we can conclude that the proposed DSIDE algorithm is more competitive with the other seven improved DE algorithms on these functions at 30 D.

From Table 3 on 100 D, the proposed DSIDE algorithm displays that 25 out of 30 benchmark functions have better or equal performance than jDE, JADE, and MPEDE, 26 out of 27 benchmark functions have better or similar performance than SaDE, CoDE, and EPSDE, and 28 out of 30 benchmark functions have better or similar performance than DEPSO. Furthermore, the proposed DSIDE algorithm performs the best on benchmark functions $f_1 \sim f_{11}$, $f_{13} \sim f_{22}$, $f_{24} \sim f_{26}$, and f_{28} in comparison with all other contrasted algorithms, performs slightly inferior on benchmark functions f_{14}, f_{23}, f_{27}, and only performs poorly on other three benchmark functions. That is to say, DSIDE has an overall better performance on benchmark functions $f_1 \sim f_{30}$ at 100 D.

From Table 4, we can see the results of Wilcoxon's rank-sum test for 30 D and 100 D problems. R^+ is the sum of positive ranks in which the first algorithm performs better than the second, and R^- is the sum of negative ranks in which the first algorithm performs worse than the second. As shown in the table, we can observe that, for all comparison of DEs, all R^+ values obtained by DSIDE are higher than R^-. It proves that DSIDE outperforms other compared DE algorithms significantly. Tables 5 and 6, respectively, utilize Friedman and Kruskal–Wallis statistical test to compare the performance of each algorithm on 30 D and 100 D problems. It can be seen that the test results obtained by DSIDE are the minimum regardless of the high dimension or low dimension, indicating that DSIDE has the best performance among the comparison algorithms.

So far, all the nonparametric tests, including Wilcoxon's rank-sum, Friedman, and Kruskal–Wallis test, support the conclusion that DSIDE is superior to other competing algorithms.

Furthermore, we compare the convergence curves of each algorithm on benchmark functions at 30 D and 100 D. All convergence curves are studied and analyzed from the aspects of convergence precision and whether they converge to the global optimum or not. Some representative convergence curves are depicted in Figures 1 and 2.

As shown in Figures 1(a) and 1(b), in convergence curves of function f_1 at 30 D and 100 D, only DSIDE converges to the global optimum, and the average convergence accuracy is much higher than other algorithms under the same generations. Convergence curves of f_7, as shown in Figures 1(c)and 1(d). Although convergence precision is not always optimal in the evolution process, only DSIDE gets the global optimum. Figures 1(e) and 1(f) show convergence curves of f_{13} at 30 D and 100 D, respectively. All algorithms have not found the optimal solution, but the average convergence accuracy of DSIDE is much higher than other algorithms under the same generations and obtains the best value. Figures 1(g) and 1(h) show convergence curves of f_{14} at 30 D and 100 D, respectively. All algorithms have not obtained the global minimum. JADE performs the best on the low-dimensional problem, while DSIDE is the best on high-dimensional. In Figures 1(i) and 1(j), DSIDE converges the fastest on f_{15}. DSIDE, EPSDE, and jDE converge to the global optimum at 30 D, while DSIDE and DEPSO reach the optimal at 100 D. In Figures 1(k) and 1(l), only DSIDE gets the global optimal

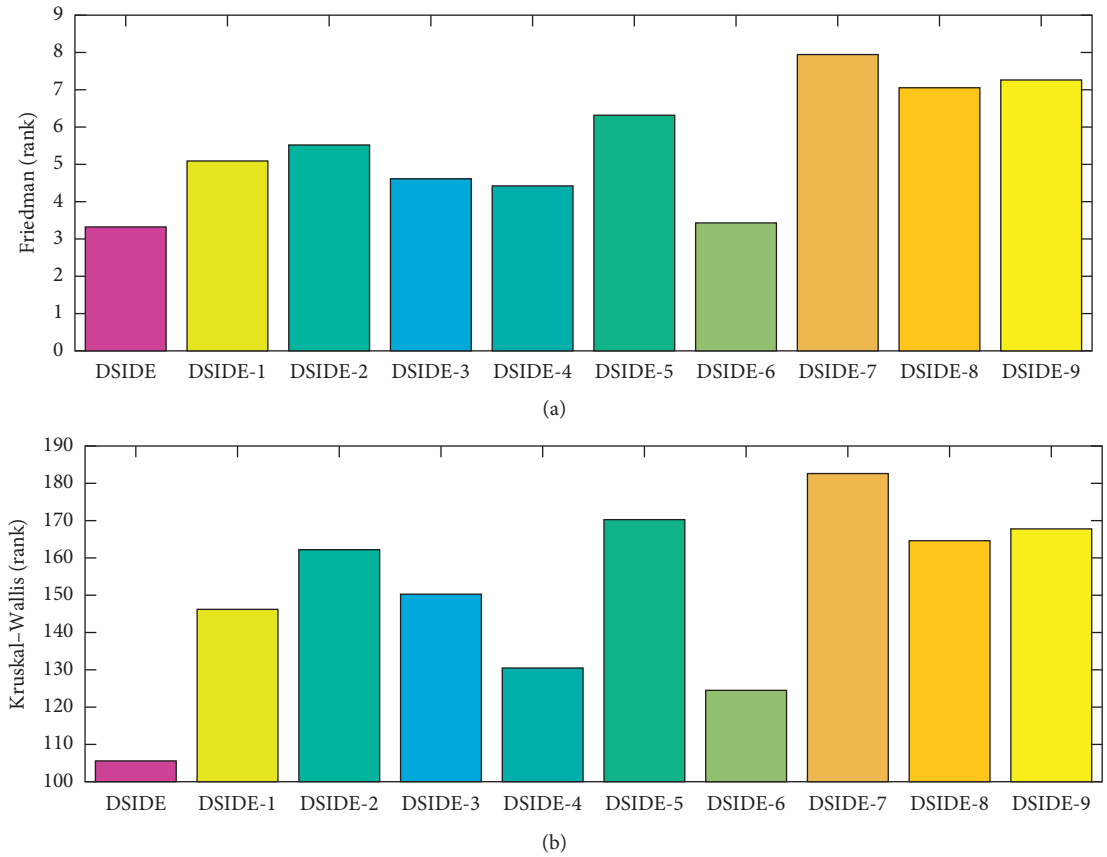


Figure 3: Nonparametric test results of proposed DSIDE and 9 DSIDE variants over 30 independent runs. (a) Friedman test results. (b) Kruskal–Wallis test results.

on f_{16} and consumes fewer generations and converges quickly.

In Figure 2(a), DSIDE, JADE, DEPSO, and EPSDE obtain the optimal on f_{18} at 30 D. In Figure 2(b), DSIDE and DEPSO get the global optimal on f_{18} at 100D. DSIDE has the fastest convergence speed in both low-dimensional and high-dimensional problems. Convergence curves of function f_{23} in Figures 2(c) and 2(d), none of the algorithms finds the global minimum, and there is a phenomenon of "evolutionary stagnation." In Figures 2(e) and 2(f) on function f_{25}, only CoDE cannot find the global minimum at 30D; DSIDE and DEPSO get the global optimal at 100D, but the former costs much less generations. In Figures 2(g) and 2(h), DSIDE converges to the global optimal on f_{26}, while other algorithms suffer from "evolutionary stagnation." In Figure 2(i), the global minimum value is found by all algorithms except CoDE and EPSDE on f_{28} at 30 D. In Figure 2(j), DSIDE and DEPSO get the global optimal on function f_{28} at 100 D. In Figures 2(k) and 2(l), DSIDE performs relatively low but consistently outperforms DEPSO on function f_{29}.

In general, through the comparative analysis of the above experiments, DSIDE not only obtains the global optimal value most times on these benchmark functions but also is superior to other algorithms in terms of convergence speed and convergence accuracy.

4.3. Efficiency Analysis of Proposed Algorithmic Components. So far, the above experiment exhibits the combined effect of the proposed DSIDE. In this section, the efficiency analysis of proposed algorithmic components is completed, including the enhanced mutation strategy of the reference factor and the adaptive strategy of the scaling factor and crossover probability. Some variants of DSIDE are listed as follows:

(i)To verify the effectiveness of the enhanced mutation strategy of reference factor α, DSIDE variants adopt dynamic F, CR, and constant reference factor of $\alpha = 0.3$ and $\alpha = 0.6$ and random real number in $[0, 1]$, which are, respectively, called as DSIDE-1, DSIDE-2, and DSIDE-3 one by one.

(ii) To investigate the validity of the scaling factor adaptive strategy, DSIDE variants employ dynamic CR, α and fixed scaling factor of $F = 0.3$, $F = 0.6$, and random real number in $[0, 1]$, which are named DSIDE -4, DSIDE -5, and DSIDE -6 for short.

(iii) To study the contribution of the crossover probability adaptive strategy, DSIDE variants with shifty F, α and settled crossover probability of CR = 0.3 and CR = 0.6, and random real number in $[0, 1]$ are, respectively, abbreviated as DSIDE-7, DSIDE-8, and DSIDE-9.

Table 7: Wilcoxon's rank-sum test results of proposed DSIDE and 9 DSIDE variants over 30 independent runs.

Comparison	$D = 30$				
	R^+	R^-	p value	$\alpha = 0.05$	$\alpha = 0.1$
DSIDE vs. DSIDE-1	216	37	$1.12e-02$	No	Yes
DSIDE vs. DSIDE-2	165	135	$5.63e-03$	Yes	Yes
DSIDE vs. DSIDE-3	154	122	$1.34e-02$	No	Yes
DSIDE vs. DSIDE-4	75	30	$2.21e-01$	No	No
DSIDE vs. DSIDE-5	194	106	$5.63e-03$	Yes	Yes
DSIDE vs. DSIDE-6	59	46	$3.00e-01$	No	No
DSIDE vs. DSIDE-7	242	34	$2.94e-03$	Yes	Yes
DSIDE vs DSIDE-8	276	0	$6.46e-03$	Yes	Yes
DSIDE vs DSIDE-9	276	0	$7.45e-03$	Yes	Yes

For the purpose of evaluating and comparing the performance of DSIDE variants, Friedman test, Kruskal–Wallis test, and Wilcoxon's rank-sum test are adopted, and the test results are shown in Figure 3(a), Figure 3(b), and Table 7, respectively. The following summaries can be obtained. (1) From Figure 3, we can observe that DSIDE and DSIDE-6 are the best and the second, while the performance of other DSIDE variants is relatively low. The combined effect of the proposed algorithmic components is the best. (2) From Table 7, the integrated DSIDE performs significantly better than DSIDE variants (DSIDE-2 and DSIDE-5) with a larger reference factor and a lager scaling factor, as well as DSIDE variants (DSIDE-7, DSIDE-8, and DSIDE-9) with different crossover probability. The performance between the integrated DSIDE and DSIDE-1 with a smaller reference factor, DSIDE-3 with a random reference factor, and DSIDE-4 with a smaller scaling factor show no significant difference when the significance level of Wilcoxon's rank-sum test is 0.1, but the difference is opposite when the significant level is 0.05. At the same time, there is no performance difference between DSIDE and DSIDE-6 with a random scaling factor, regardless of the significance level. The validity of the proposed mutation strategy and adaptive strategy for control parameters is demonstrated utilizing above experimental comparisons. It is noted that the contribution of the adaptive strategy of crossover probability is larger than enhanced mutation strategy and adaptive strategy of scaling factor. That is to say, although the enhanced mutation strategy of reference factor and adaptive strategy of scaling factor are effective, DSIDE is less susceptible to both a smaller or variational reference factor and scaling factor.

5. Conclusions

DSIDE's innovation lies in two strategies, the enhanced mutation strategy and the novel adaptive strategy for control parameters. On the one hand, the enhanced mutation strategy considers the influence of the reference individual on the overall evolution. It introduces the reference factor, which is beneficial to global exploration in the early stage of evolution and global convergence in the later stage. On the other hand, the novel adaptive strategy for control parameters can dynamically adjust the scaling factor and crossover probability according to the fitness value, which has a positive impact on maintaining the population diversity. DSIDE is compared with other seven DE algorithms, the results are evaluated by three nonparametric statistical tests, and the convergence curves are analyzed. Experimental results show that the proposed DSIDE can effectively improve the optimization performance. Besides, the efficiency analysis of proposed algorithmic components has been carried out, which further proves the comprehensive effect and validity of DSIDE.

So far, DE variants have been applied to various fields, such as target allocation [41], text classification [42], image segmentation [43], and neural network [44–47]. For the future work, the proposed DSIDE algorithm will be applied to the parameter optimization of neural network and may further apply it to the air traffic control system for flight trajectory prediction [48, 49].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, no. U1833128.

References

[1] R. Storn and K. Price, "Differential evolution: a simple and efficient adaptive scheme for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 23, no. 1, 1995.

[2] A. E. Eiben, R. Hinterding, and Z. Michalewicz, "Parameter control in evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 2, pp. 124–141, 1999.

[3] G. Karafotias, M. Hoogendoorn, and A. E. Eiben, "Parameter control in evolutionary algorithms: trends and challenges," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 2, pp. 167–187, 2015.

[4] M. G. H. Omran, A. Salman, and A. P. Engelbrecht, "Self-adaptive differential evolution," in *Proceedings of the International Conference on Computational Intelligence and Security*, Xi'an, China, 2005.

[5] J. Liu and J. Lampinen, "A fuzzy adaptive differential evolution algorithm," *Soft Computing*, vol. 9, no. 6, pp. 448–462, 2005.

[6] J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer, "Self-adapting control parameters in differential evolution: a comparative study on numerical benchmark problems," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 6, pp. 646–657, 2006.

[7] N. Noman, D. Bollegala, and H. Iba, "An adaptive differential evolution algorithm," in *Proceedings of the IEEE Congress of Evolutionary Computation*, IEEE, New Orleans, LA, USA, 2011.

[8] M. Asafuddoula, T. Ray, and R. Sarker, "An adaptive hybrid differential evolution algorithm for single objective optimization," *Applied Mathematics and Computation*, vol. 231, pp. 601–618, 2014.

[9] R. Tanabe and A. Fukunaga, "Success-history based parameter adaptation for differential evolution," in *Proceedings of the 2013 IEEE Congress on Evolutionary Computation*, pp. 71–78, IEEE, Cancun, Mexico, June 2013.

[10] R. Tanabe and A. S. Fukunaga, "Improving the search performance of shade using linear population size reduction," in *Proceedings of the 2014 IEEE Congress on Evolutionary Computation*, pp. 1658–1665, IEEE, Beijing, China, 2014.

[11] W. Zhu, Y. Tang, J.-A. Fang, and W. Zhang, "Adaptive population tuning scheme for differential evolution," *Information Sciences*, vol. 223, pp. 164–191, 2013.

[12] S. Zhao, X. Wang, L. Chen, and W. Zhu, "A novel self-adaptive differential evolution algorithm with population size adjustment scheme," *Arabian Journal for Science and Engineering*, vol. 39, no. 8, pp. 6149–6174, 2014.

[13] J. S. Pan, C. Yang, F. J. Meng, Y. X. Chen, and Z. Y. Meng, "A parameter adaptive DE algorithm on real-parameter optimization," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 1, pp. 1–12, 2020.

[14] Z. Meng, Y. Chen, and X. Li, "Enhancing differential evolution with novel parameter control," *IEEE Access*, vol. 8, pp. 51145–51167, 2020.

[15] M. Di Carlo, M. Vasile, and E. Minisci, "Adaptive multi-population inflationary differential evolution," *Soft Computing*, vol. 24, no. 5, pp. 3861–3891, 2020.

[16] D. Wales and J. Doye, "Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms," *Journal of Physical Chemistry A*, vol. 101, no. 28, pp. 5111–5116, 1998.

[17] S. Li, Q. Gu, W. Gong, and B. Ning, "An enhanced adaptive differential evolution algorithm for parameter extraction of photovoltaic models," *Energy Conversion and Management*, vol. 205, Article ID 112443, 2020.

[18] S. Das, A. Abraham, U. K. Chakraborty, and A. Konar, "Differential evolution using a neighborhood-based mutation operator," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 3, pp. 526–553, 2009.

[19] J. Q. Zhang and A. C. Sanderson, "JADE: adaptive differential evolution with optional external archive," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 945–958, 2009.

[20] A. K. Qin, V. L. Huang, and P. N. Suganthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 2, pp. 398–417, 2009.

[21] Y. Wang, Z. Cai, and Q. Zhang, "Differential evolution with composite trial vector generation strategies and control parameters," *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 1, pp. 55–66, 2011.

[22] M. G. Epitropakis, D. K. Tasoulis, N. G. Pavlidis, V. P. Plagianakos, and M. N. Vrahatis, "Enhancing differential evolution utilizing proximity-based mutation operators," *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 1, pp. 99–119, 2011.

[23] R. Mallipeddi, P. N. Suganthan, Q. K. Pan, and M. F. Tasgetiren, "Differential evolution algorithm with ensemble of parameters and mutation strategies," *Applied Soft Computing*, vol. 11, no. 2, pp. 1679–1696, 2011.

[24] W. L. Xiang, X. L. Meng, M. Q. An, Y. Z. Li, and M. X. Gao, "An enhanced differential evolution algorithm based on multiple mutation strategies," *Computational Intelligence and Neuroscience*, vol. 2015, Article ID 285730, 15 pages, 2015.

[25] L. Cui, G. Li, Q. Lin, J. Chen, and N. Lu, "Adaptive differential evolution algorithm with novel mutation strategies in multiple sub-populations," *Computers & Operations Research*, vol. 67, pp. 155–173, 2016.

[26] G. Wu, R. Mallipeddi, P. N. Suganthan, R. Wang, and H. Chen, "Differential evolution with multi-population based ensemble of mutation strategies," *Information Sciences*, vol. 329, pp. 329–345, 2016.

[27] Z. Meng, J.-S. Pan, and L. Kong, "Parameters with adaptive learning mechanism (palm) for the enhancement of differential evolution," *Knowledge-Based Systems*, vol. 141, pp. 92–112, 2018.

[28] Z. Meng, J.-S. Pan, and K.-K. Tseng, "PaDE: an enhanced differential evolution algorithm with novel control parameter adaptation schemes for numerical optimization," *Knowledge-Based Systems*, vol. 168, pp. 80–99, 2019.

[29] Z. Wei, X. Xie, T. Bao, and Y. Yu, "A random perturbation modified differential evolution algorithm for unconstrained optimization problems," *Soft Computing*, vol. 23, no. 15, pp. 6307–6321, 2019.

[30] M. Duan, H. Yang, H. Liu, and J. Chen, "A differential evolution algorithm with dual preferred learning mutation," *Applied Intelligence*, vol. 49, no. 2, pp. 605–627, 2019.

[31] M. Tian and X. Gao, "Differential evolution with neighborhood-based adaptive evolution mechanism for numerical optimization," *Information Sciences*, vol. 478, pp. 422–448, 2019.

[32] S. H. Wang, Y. Z. Li, and H. Y. Yang, "Self-adaptive mutation differential evolution algorithm based on particle swarm optimization," *Applied Soft Computing*, vol. 81, 2019.

[33] Z. Meng and J.-S. Pan, "HARD-DE: hierarchical archive based mutation strategy with depth information of evolution for the enhancement of differential evolution on numerical optimization," *IEEE Access*, vol. 7, pp. 12832–12854, 2019.

[34] L. M. Zheng, S. X. Zhang, K. S. Tang, and S. Y. Zheng, "Differential evolution powered by collective information," *Information Sciences*, vol. 399, pp. 13–29, 2017.

[35] J.-S. Pan, N. Liu, and S.-C. Chu, "A hybrid differential evolution algorithm and its application in unmanned combat aerial vehicle path planning," *IEEE Access*, vol. 8, pp. 17691–17712, 2020.

[36] Z. Meng, C. Yang, X. Li, and Y. Chen, "Di-DE: depth information-based differential evolution with adaptive parameter control for numerical optimization," *IEEE Access*, vol. 8, pp. 40809–40827, 2020.

[37] J. J. Liang, B. Y. Qu, P. N. Suganthan, and Q. Chen, "Problem definition and evaluation criteria for the cec 2015 competition on learning based real-parameter single objective optimization," 2015.

[38] N. H. Awad, M. Z. Ali, P. N. Suganthan, J. J. Liang, and B. Y. Qu, "Problem definitions and evaluation criteria for the CEC 2017 special session and competition on single objective real-Parameter numerical optimization," 2016.

[39] P. M. Bilal, M. Pant, H. Zaheer, L. Garcia-Hernandez, and A. Abraham, "Differential evolution: a review of more than two decades of research," *Engineering Applications of Artificial Intelligence*, vol. 90, p. 103479, 2020.

[40] J. Derrac, S. García, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3–18, 2011.

[41] A. E. Bayrak and F. Polat, "Employment of an evolutionary heuristic to solve the target allocation problem efficiently," *Information Sciences*, vol. 222, no. 3, pp. 675–695, 2013.

[42] A. Onan, S. Korukoğlu, and H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Systems with Applications*, vol. 62, pp. 1–16, 2016.

[43] O. Tarkhaneh and H. F. Shen, "An adaptive differential evolution algorithm to optimal multi-level thresholding for MRI brain image segmentation," *Expert Systems with Applications*, vol. 138, 2019.

[44] B. Subudhi and D. Jena, "A differential evolution based neural network approach to nonlinear system identification," *Applied Soft Computing*, vol. 11, no. 1, pp. 861–871, 2011.

[45] H. Dhahri, A. M. Alimi, and A. Abraham, "Hierarchical multi-dimensional differential evolution for the design of beta basis function neural network," *Neurocomputing*, vol. 97, pp. 131–140, 2012.

[46] H.-C. Lu, M.-H. Chang, and C.-H. Tsai, "Parameter estimation of fuzzy neural network controller based on a modified differential evolution," *Neurocomputing*, vol. 89, pp. 178–192, 2012.

[47] A. P. Piotrowski, "Differential Evolution algorithms applied to neural network training suffer from stagnation," *Applied Soft Computing*, vol. 21, pp. 382–406, 2014.

[48] S. Hong and K. Lee, "Trajectory prediction for vectored area navigation arrivals," *Journal of Aerospace Information Systems*, vol. 12, no. 7, pp. 490–502, 2015.

[49] C. E. V. Gallego, V. F. Gomez, M. A. A. Carmona, R. M. A. Valdes, F. J. S. Nieto, and M. G. Martinez, "A machine learning approach to air traffic interdependency modelling and its application to trajectory prediction," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 356–386, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9287241, 18 pages
https://doi.org/10.1155/2020/9287241

Research Article

Investigation of an Underwater Vectored Thruster Based on 3RPS Parallel Manipulator

Tao Liu,[1] Yuli Hu,[2] and Hui Xu[1]

[1]*School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China*
[2]*Key Laboratory for Unmanned Underwater Vehicle, Northwestern Polytechnical University, Xi'an 710072, China*

Correspondence should be addressed to Yuli Hu; yulihu001@gmail.com and Hui Xu; merleliu@mail.nwpu.edu.cn

Received 24 July 2020; Revised 5 September 2020; Accepted 11 September 2020; Published 29 September 2020

Academic Editor: Mohamed El Ghami

Autonomous underwater vehicles (AUVs) are important and useful tool platforms in exploring and utilizing ocean resource. However, the effect of control surfaces would decrease even invalid complete in this condition, and it is very hard for conventional AUVs to perform detailed missions at a low forward speed. Therefore, solving this problem of AUVs becomes particularly important to increase the application scope of AUVs. In this paper, we present a design scheme for the vectored thruster AUV based on 3RPS parallel manipulator, which is a kind of parallel manipulator and has advantages of compact structure and reliable performance. To study the performance and characteristics of the proposed thrust-vectoring mechanism, a series of works about corresponding kinematic and dynamic analysis have been performed through the theoretical analysis and numerical simulation. In the part of kinematics, the inverse, forward kinematics, and workspace analysis of the thrust-vectoring mechanism is presented, and the numerical simulations are accomplished to prove the feasibility and effectiveness of this design in AUVs. In order to further verify feasibility of the thrust-vectoring mechanism, based on the considerations of various affecting factors, a dynamic model of the designed thrust-vectoring mechanism is established according to theoretical analysis, and the driving forces of the linear actuator are presented through a series of numerical simulations. In addition, a control scheme based on PID algorithm is proposed for the designed vectored thruster with considering various affecting factors and the application environment. Meanwhile, the control scheme is also established and verified in MATLAB Simscape Mutibody. A series of numerical simulations of the thrust-vectoring mechanism prove the feasibility of the vectored thruster. According to equipping the designed vectored thruster, the AUVs can overcome the limit of weakening the control ability at zero or low forward speeds, and this improvement also expands the application of it, which has been scaled greatly.

1. Introduction

Over the last few decades, due to exhausting of resources and energy, human beings are bearing with a series of survival predicaments and development challenges [1, 2]. Because of the lack of land-based resource and the continuing need for all kinds of resources, an increasing number of countries and scientists have paid more and more attention on the exploitation and utilization of resources [3–5]. In the present, most of the water available on Earth exists in the oceans, yet only a small part of this vast resource has been explored [2]. The ocean has vast areas and is rich in all kinds of natural resources, such as marine life, oil, natural gas, and minerals. Additionally, the ocean not only contains a lot of marine recourse but also brings a lot of traffic convenience around the world. With the progress of society and economy, the development of mineral resource has become an inevitable trend. Exploring and exploiting the oceans has become the principal development strategy of every country in the world. However, the nature environment of the ocean is too harsh to explore, and the advanced technology has been rapidly developed in recent years, such as autonomous underwater vehicles (AUVs), remotely operated vehicles (ROVs) [6–8], and unmanned marine vehicles (USVs) [9]. In addition, other different techniques have also been used for controlling all kinds of robots, such as proportional integral derivate [10], fuzzy control [11, 12], and sliding model control [13].

AUVs have become a main tool for surveying below the sea due to the great improvement in their performance and advancement in underwater research. Through equipping a large quantity of advanced instruments and equipment, AUVs are capable of accomplishing applications including scientific, commercial, and military tasks such as exploration of oceans [14, 15], oceanography mapping surveys [16–18], the collecting ocean environment information [19–21], and searching and rescuing for shipwrecks [22, 23] and debris from the missing airplanes [24, 25]. With the expanding area of applications, the design of AUVs needs to meet the higher demand continuously. Although it has made great steps in AUVs performance, the new emerging technologies and demands for exploiting oceans have attracted critical mass of scientists and engineers to undertake the research of AUVs.

Conventional AUVs are designed equipped with a main propeller and control surfaces at the tail cone for propulsion and control [26–28]. These conventional AUVs can fulfill the work well under normal conditions. When the conventional AUVs need to complete exploration tasks with a lower speed in a complex and unknown underwater environment, the control capability of AUV depends heavily on the control faces made up of fin and rudder. The velocity of AUV is relatively low or zero because of the demand of practical problems, such as scanning topographic map, taking photographs, and monitoring marine observation data. However, they are unable to perform detailed inspection missions at zero or low forward speeds due to the control faces which become ineffective in this condition [28, 29]. The cause of this problem is that the generation of control forces from control surfaces depends on forward speeds of AUVs [28, 30]. Therefore, this disadvantage has further development and application of conventional AUVs greatly.

There are some approaches to solve this problem, such as installing additional thrusters to provide additional control forces for controlling AUVs [28, 31–34], but this method would result in the problems of complicated structure and increasing energy consumption. Its complex structure, adverse working environment, and so on causes the decrease of reliability of the whole AUV control system. Installation and maintenance of additional thrusters would significantly increase energy expenditure or energy carrier for sailing.

Another more efficient and workable method to release this restriction is to use vectored thruster to replace the conventional propulsion types [7, 29, 35–38]. The AUVs equipped with vectored thruster do not require the use of fin and rudder for controlling at all. Since this kind of AUVs driven by a vectored thruster, the control forces are generalized force components produced by vectored thruster, and these forces only depend on the rotational speed of the propeller. Therefore, the AUVs equipped with vectored thruster are independent from any control forces generated by control faces, and the controllability of vehicle is markedly improved and obtains a better good application effect. So, the vectored thruster AUVs are capable of accomplishing detailed missions at a low forward speed.

In the research areas of this field, some companies and research institutions have made progress in theory study and application of vectored thrusters [29, 36, 38]. Among above research AUVs, Bluefin and MBARI have achieved great successes and provided considerable experience in the use and study of vectored thrusters. More importantly, the engineering practice of Bluefin and MBARI shows that this method can raise control efficiency greatly and also reduce the possibility of losing control at low speed. However, the existing design of the vectored thruster is almost designed based on serial mechanism; this kind of mechanism has the disadvantages of complex structure, low bearing capacity, and high moment inertia. So, based heavily on practicalities of serial mechanism, the existing design of the vectored thruster is too bulky and complex to use for AUVs. Considering the restrictions of application environment and structure size, it is crucial to choose a suitable mechanism for designing new vectored thruster AUVs.

On the contrary, compared with other commonly used mechanical structures, parallel manipulators have numerous advantages, such as small size, compact and reasonable structure, reliable performance, fast response, high positioning precision, high stability, high sensitivity, high stiffness, and better dynamic performance [39–42]. Those merits of parallel manipulators make the device have high popularization value and use value, such as medical and industrial robots, flight simulator, and mechanical device. Inspired by various applications of parallel manipulators, the idea of vectored thruster AUVs based on parallel manipulator is generated. In the field of thrust-vectoring mechanism research, many scientists and engineers have made great contributions to the development of vectored thruster AUVs based on parallel manipulators. The full deflection vectored thruster is based on the spatial linkage and universal joint proposed by Cavallo and Michelini [43]; the authors designed a 3-SPS-S parallel manipulator with passive constraining spherical joints to drive the underwater vehicle [30]. The above thrusters currently have some problems, such as the structure is relatively complex, and the motion real-time resolving method and the dynamics model for the vectored thruster are difficult.

With comparing structure characteristics of different kinds of parallel manipulators and considering actors of application environment, 3RPS is chosen from various parallel manipulators as the thrust-vectoring mechanism mainly various advantages, including its compact structure, high position tracking precision, and fast response speed. This parallel mechanism is a strong coupled nonlinear structure, so its motion control is too complex to use more widely [44–47]. Despite its advantages, the 3RPS parallel manipulator also needs to overcome some problems that would restrict the development and application of the thrust-vectoring mechanism. Through reading and analyzing the domestic and foreign related literature, various methods of kinematics and dynamics for 3RPS parallel manipulators have been presented [44, 45, 48–50].

On the basis of the above considerations, the design concept of vectored thruster which is made up of 3RPS parallel manipulators is introduced. The vectored thruster based on 3RPS parallel manipulators has terse structure, convenient operation, convenient installation, steady working system, and wide adjustable range. Using this

method, the AUVs are able to provide the vectored thrust effectively and efficiently. More than anything, the AUVs equipped with vectored thruster are able to complete a variety of the complex tasks at a comparably low forward speed.

In this paper, the structural design of the vectored thruster based on 3RPS is introduced briefly. In order to satisfy the design requirements and study the motion characteristics of vectored thruster, the kinematics and dynamics model of the thrust-vectoring mechanism are established, and the related simulation is presented to verify feasibility of the scheme. Finally, a control scheme for the vectored thruster is designed and simulated in Matlab. The theoretical analysis and numerical simulations prove that the proposed vectored thruster based on a 3RPS parallel manipulator can effectively realize the function of providing the required vectored thrust for thrust-vectoring propulsion.

2. The General Design of Vectored Thruster

The configuration of the whole AUV equipped with the designed vectored thruster based on 3RPS parallel manipulator is presented, as shown in Figure 1. Due to the existence of the vectored thruster, the AUVs do not need any more rudders to provide control forces. The force generated by the vectored thruster is used as control force for AUV's yaw and pitch motion. Consequently, the tilt angle is important one of the criteria to assessing the performance of the proposed vectored thruster. However, the space in the stern of AUVs is limited, and the tilt angle range of the designed vectored thruster is also limited. Referring to a literature review [29, 51], the duct propeller's tilt angle is limited to plus or minus 15° in our design. In addition, the vectored thruster contains a duct that can be used for protecting the propeller form damaging and enhancing flow capability.

In terms of structure design, considering the specific requirements of application environment and the stability of system, we adopted the modular design for vectored thruster AUV. The designed vectored thruster is mounted on the stern of an AUV as an integral and independent, which is adopted for convenient installation and maintenance. The designed vectored thrust duct propeller system mainly contains the duct propeller and the thrust-vectoring mechanism. A whole structure model of the vectored thruster AUV based on 3RPS parallel manipulator is built up, as shown in Figure 2.

At present, the duct propeller is the most widely used form of propulsion device for underwater robots. A duct propeller is mainly composed of an annular wing and a propeller. There are many underwater vehicles equipped with duct propellers, for the extraordinary performance of improving the propulsive efficiency and avoiding cavitation conditions [52]. This kind of propeller is able to provide the thruster from zero to cruising speed more effectively. Just because of an effectively increased thrust in the condition of a low forward speed, the duct propeller is widely used in various marine vessels, such as AUVs and ROVs.

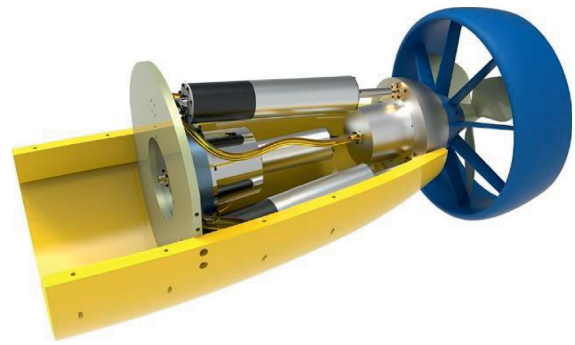


Figure 1: Vectored thruster mounted at AUV.

Furthermore, underwater environments are very complex and harsh; propellers are very easily destroyed by underwater animals and plants, waves, even currents, and other uncertainties. Hence, the existence of a duct can protect the propeller against damage from the underwater environment during all kinds of missions. Moreover, since the duct also can generate the thrust during the voyage, the duct is an important source of control force for AUV's yaw and pitch motion.

In our design, the duct propeller is driven by a main electric motor installed in the rotating body, which is aligned with the holes of the duct's inner shaft with fastening screws. In order to simplify the unnecessary transmission structure and reduce the redundant weight, the main motor has been bedded on the rotating body with a duct propeller. It is very clean and efficient to take direct connects with the propeller and the main motor. This installation mode of the main motor and rotating body can improve space utilization significantly and reduce the weight of vectored thruster effectively. And this modularity makes the duct propeller system easy to maintain and debug on the whole vectored thruster control system.

As the implementing actuator of the vectored thruster, the thrust-vectoring mechanism is fundamental to the overall system for its basic functions. There are many methods on how to realize thrust vectoring, and each method has its own advantages and disadvantages. Considering the limited space of AUVs' tailcone and the harsh operation condition, it is central to choose an appropriate mechanism structure that can complete the design function of achieving vector control effectively for AUVs. Comparing to the serial mechanism, parallel manipulators have many inherent superiorities, such as small size, compact and reasonable structure, reliable performance, fast response, high positioning precision, high stability, high sensitivity, high stiffness, and better dynamic performance.

Integrating practical application environment of AUVs and based on the application background of various parallel manipulators, 3RPS parallel manipulator is chosen as the thrust-vectoring mechanism after analyzing various mechanical structures. In accordance with this notion, a novel thrust-vectoring mechanism based on the 3RPS parallel manipulator for AUVs is designed, as shown in Figure 3.

The thrust-vectoring mechanism is designed based on 3-RPS manipulator, which has a top rotting platform, a fixed base, and three identical sets of driving limbs and joints.

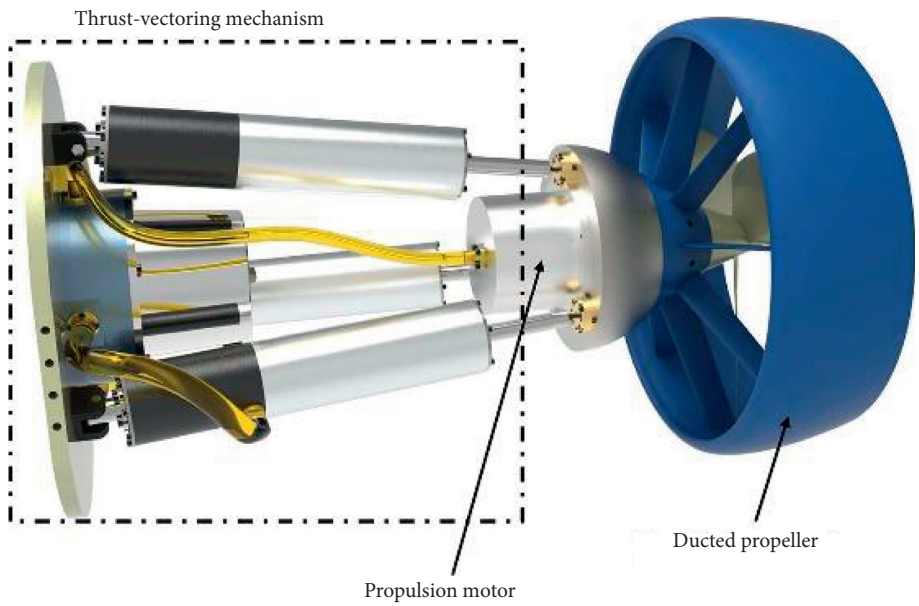


Figure 2: Vectored duct thrust propeller.

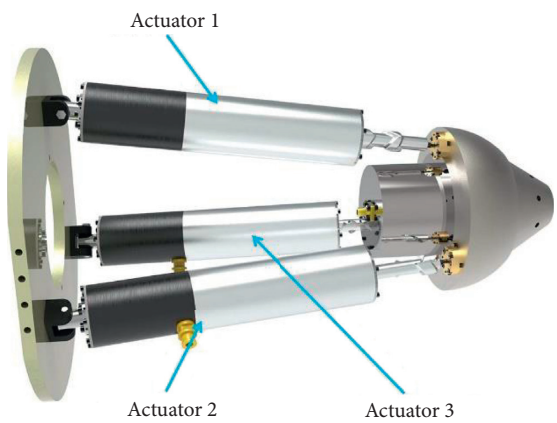


Figure 3: Thrust-vectoring mechanism based on 3-RPS parallel manipulator.

Each driving limb has an actuating prismatic joint (P) attached to the fixed base by a rotational joint (R) and connected to the platform by a spherical joint (S) [53–55]. Each limb is driven by a linear actuator. Thus, the length of limb could be changed within the operation range, and the top platform would rotate when the length of limbs changes with a certain law of motion. The vectored thruster is installed on the tail horizontally, which connected with the shell of the AUV via fastening screws. The duct propeller can rotate around the center of the top platform horizontally and vertically, and the thrust generated by the vectored thruster will drive the vehicle forward or changes the direction of movement.

The 3RPS parallel manipulator has two rotational and one translational degree of freedom (DOF). It is superfluous to have the translational DOF for the thrust-vectoring mechanism, the redundant DOF needs to be constrained by motion control, and the other two rotational DOFs are used to realize functions of the thrust-vectoring mechanism. In addition, the translational DOF of 3RPS parallel manipulator will make the top rotating body bump into the shell of the AUV. So, the importance of the study on redundant DOF of the 3RPS parallel manipulator is obvious for the actual application of the thrust-vectoring mechanism.

Since the vectored thruster could generate required control forces for controlling AUVs motion, there is no need to have extra rudders as conventional AUVs. The component of the thrust as control forces is dependent on the deflection angle and the thrust of the vectored thruster. Therefore, the research on deflection angle of the vectored thruster is essential for controlling the motions of AUVs. However, it is very difficult to measure the tilt angle of the vectored thruster directly because the limited space and underwater environment is not suitable for installing sensors to measure. Another common and efficient approach to get the rotation angles is using the kinematic analysis method, which can obtain the tilt angle by measuring the lengths of the three limbs. Based on this kinematic method, tilt angle information about the vectored thruster can be obtained via relative calculation with the lengths of three limbs, which can be measured directly by length sensors installed in actuators.

In order to realize precision and stable positioning control of the proposed vectored thruster, the design of the automation control system is fundamental to achieve objectives. Hence, establishment of kinematic and dynamic models for the thrust-vectoring mechanism based on the 3RPS parallel manipulator is significant to achieving perfect control of the vectored thruster based on the above analysis.

3. Kinematic Analysis of the Thrust-Vectoring Mechanism

The thrust-vectoring mechanism is designed based on the 3-RPS parallel manipulator, which is composed of a base plate, a rotating platform, and three uniformly distributed driving

limbs, as shown in Figure 3. According to the needs of analyzing the motion of the top rotating platform, two Cartesian coordinate systems with associated symbols have been established in the 3RPS parallel manipulator and shown in Figure 4. The reference frame *O-xyz*, which is the global coordinate system, is fixed to the center of the immovable base and the *z*-axis normal to the fixed base. Similarly, the reference frame *P-ijk* denotes the local coordinate system located on center point *P* of the rotating platform, whose *j*-axis is perpendicular to the bottom surface of the platform.

The moving sides of driving limbs (linear actuators) are connected to the upper rotating platform through three sphere joints that is fixed directly to the center of the top platform, while the other sides of the limbs are connected to the base with three revolute joints that are symmetrical about the center of base. A_1, A_2, and A_3 are the connected points between the fixed base and the driving legs (linear actuators), B_1, B_2, and B_3 denote the points of the revolute joints. It should be mentioned that A_1B_1, A_2B_2, and A_3B_3 are perpendicular to the fixed base because A_1B_1, A_2B_2, and A_3B_3 represent revolute joints with a certain height. A reference frame O'-*xyz* is established with respect to the plane formed by points B_1, B_2, and B_3, and this plane parallels with the fixed base A_1, A_2, and A_3. The connected points between the moving parts and the rotating platform are expressed as C_1, C_2, and C_3. The radius of the fixed base and the top platform are defined as r_1 and r_2, *O* and *P* denote the center points of the base and top platform, respectively. L_1, L_2, and L_3 denote the lengths of three linear actuators between the top platform and the fixed base.

As we can see in Figure 4, in the global reference frame *O-xyz*, the center point of the equilateral triangle made up of three points A_1, A_2, and A_3 is expressed as *O*, and the radius of the fixed base is defined as $OA_1 = OA_2 = OA_3 = r_1$. Hence, the location of A_i in global reference frame *O-xyz* can be described as follows:

$$P_{A-O} = \left[\, A_1 \;\; A_2 \;\; A_3 \,\right] = \begin{bmatrix} 0 & r_1 & -r_1 \\ -\dfrac{\sqrt{3}\, r_1}{2} & \dfrac{r_1}{2} & \dfrac{r_1}{2} \\ 0 & 0 & 0 \end{bmatrix}. \tag{1}$$

Similarly, B_1, B_2, and B_3 denote the axes of revolution of the revolute joints with a certain height h_r, and this plane parallel is with the fixed base. Hence, the locations of point B_i can be denoted as follows:

$$P_{B-O} = \left[\, B_1 \;\; B_2 \;\; B_3 \,\right] = \begin{bmatrix} 0 & r_1 & -r_1 \\ -\dfrac{\sqrt{3}\, r_1}{2} & \dfrac{r_1}{2} & \dfrac{r_1}{2} \\ h_r & h_r & h_r \end{bmatrix}. \tag{2}$$

A local coordinate system *P-ijk* is established on the rotating platform bottom surface of the thrust-vectoring mechanism, in which the origin point *P* is the circumcenter of triangle C_1, C_2, and C_3. So, the locations of connection point between the linear actuators and the top platform can be described as follows:

$$P_{C-P} = \left[\, B_1 \;\; B_2 \;\; B_3 \,\right] = \begin{bmatrix} 0 & r_2 & -r_2 \\ -\dfrac{\sqrt{3}\, r_2}{2} & \dfrac{r_2}{2} & \dfrac{r_2}{2} \\ 0 & 0 & 0 \end{bmatrix}. \tag{3}$$

From Figure 4, *p* as a position vector denotes the translation vector from the center point *O* to point *P* of top rotating platform in global reference frame *O-xyz*. To describe the relative motion between top rotating platform and fixed base, a rotation matrix about frame *P-ijk* with respect to the fixed base reference frame *O-xyz* needs to be established. The position vector *p* and the rotation matrix *R* are defined as follows:

$$\begin{aligned} p &= \left[\, P_x \;\; P_y \;\; P_z \,\right]^T, \\ R &= \begin{bmatrix} c\gamma c\beta & c\gamma s\beta s\alpha - s\gamma c\alpha & c\gamma s\beta c\alpha + s\gamma s\alpha \\ s\gamma c\beta & s\gamma s\beta s\alpha + c\gamma c\alpha & s\gamma s\beta c\alpha - c\gamma s\alpha \\ -s\beta & c\beta s\alpha & c\beta c\alpha \end{bmatrix}, \end{aligned} \tag{4}$$

where $s\,(\cdot) = \sin\,(\cdot)$, $c\,(\cdot) = \cos\,(\cdot)$, and α, β, and γ denote the rotation angles about the *k*-axis, *j*-axis, and *i*-axis, respectively.

The thrust-vectoring mechanism only needs two rotational DOFs to realize its functionality; the 3RPS parallel manipulator has one more translational DOF that is redundant. In order to present the condition of the top rotating platform of the thrust-vectoring mechanism, the rotation angles are also important parameters that need to be defined. According to the need of the thrust-vectoring mechanism, the displacement between centers of the base and the top platform are set as *h*. A generalized vector *q* is established to describe the position and orientation of the top rotating platform in the global reference frame as follows:

$$q = \left[\, P_x \;\; P_y \;\; P_z \;\; \alpha \;\; \beta \;\; \gamma \,\right]^T, \tag{5}$$

where P_x, P_y, and γ are associated with the rotation angles α and β. Based on $P_x = r_2\,(s\alpha s\beta c\gamma - c\alpha s\gamma)$, $P_x = r_2\,(s\alpha s\beta c\gamma - c\alpha s\gamma)$, $P_x = r_2\,(s\alpha s\beta c\gamma - c\alpha s\gamma)$, $P_x = r_2\,(s\alpha s\beta c\gamma - c\alpha s\gamma)$, $P_y = (r_2/2)\,[(c\beta c\gamma - s\alpha s\beta s\gamma) - c\alpha c\gamma]$, $L_P = \sqrt{P_x^2 + P_y^2}$, and $\gamma = \tan^{-1}\,[s\beta s\alpha/(s\beta + c\alpha)]$, so P_x and P_y also can be written as follows:

$$\begin{cases} P_x = r_2 \cdot \delta_{Px}, \\ P_y = r_2 \cdot \delta_{Py}, \\ L_P = r_2 \cdot \delta_{LP}, \end{cases} \tag{6}$$

where δ_{Px}, δ_{Py}, and δ_{Ly} denote the influence factor of P_x, P_y, and L_P. According to the application needs to be designed in

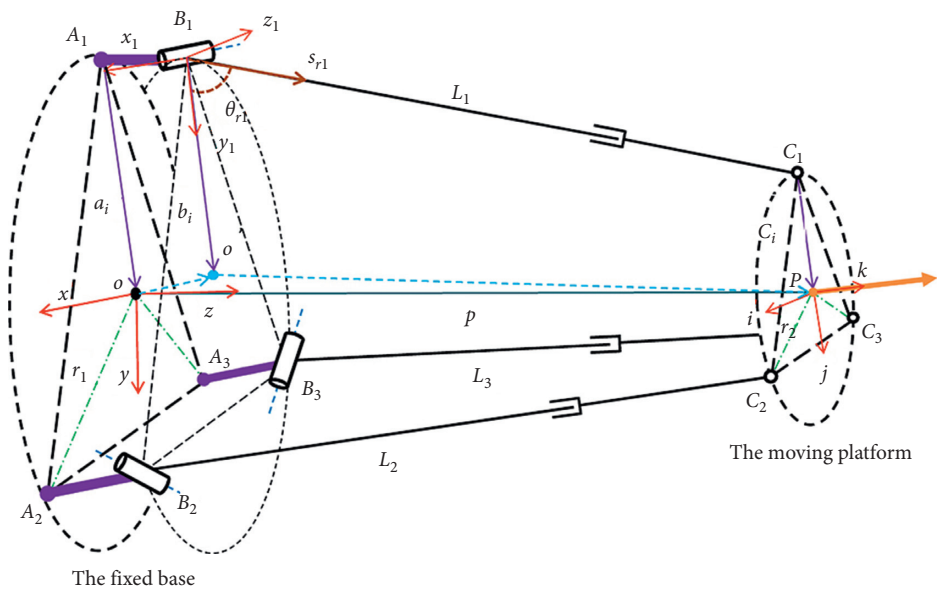


Figure 4: Schematic diagram of the 3-RPS parallel manipulator.

this design, the numerical simulations about δ_{Px}, δ_{Py}, and δ_{LP} and the angle of rotation of γ can be calculated, and the simulation results are obtained and plotted in Figure 5.

From Sections 3.1 and 3.3, the inverse, forward kinematic analyses, and workspace analysis of the thrust-vectoring mechanism are performed, and numerical simulations are conducted to validate the accuracy and reliability accordingly.

3.1. Inverse Position Analysis of Thrust-Vectoring Mechanism. In this design, inverse position analysis of the thrust-vectoring mechanism is carried out to establish the mapping relations between the position and orientation of top moving platform and the lengths of three driving linear actuators. Referring to Figure 4, the length of linear actuator L_i with respect to the fixed base reference frame O-xyz can be written as

$$L_i = \left|Rc_i + p - a_i\right|, \tag{7}$$

where a_i (c_i) denotes the vector from point O (P) to point A_i (C_i) in frame O-xyz (P-ijk).

The length change of the ith limb can be calculated as

$$\Delta L_i = L_i - L_{\text{ave}}, \tag{8}$$

where L_{ave} is the initial length of the linear actuators at the tilt angle $\alpha = \beta = 0^\circ$.

According to (7) and (8), the length changes of the three linear actuators can be obtained with related parameters presented in Table 1, and the dimension parameters of vectored thruster are calculated through CAD software. The results of the length change of ith linear actuator are plotted in Figure 6.

To further study the relationship between the tilt angles of the top rotating platform and the lengths of three linear actuators, the top moving platform moves according to $\alpha_s = \pi/9 \cdot \sin(t)$ rad and $\beta_s = \pi/9 \cdot \cos(t)$ rad. When the top rotating platform moves according to $\alpha = \alpha_s$ and $\beta = \beta_s$, based on the kinematic analysis of the thrust-vectoring mechanism above, the length of linear actuators is plotted in Figure 7.

3.2. Forward Position Analysis of Thrust-Vectoring Mechanism. Similarly, the forward position analysis of the thrust-vectoring mechanism is established to study the mapping relations between the lengths, three linear actuators, and the position and orientation of top moving platform. The position and orientation of the top moving platform is obtained according to the given length of the ith linear actuator.

Referring to Figure 4, the position vector of point C_i on the top rotating platform in global frame O-xyz can be expressed as

$$OC_i = OA_i + A_iB_i + B_iC_i, \tag{9}$$

where B_iC_i denotes the vector of the ith linear actuator, which can be expressed as

$$B_iC_i = \begin{bmatrix} L_i\, c(\theta_{1i}) c(\lambda_i) \\ L_i c(\theta_{1i}) s(\lambda_i) \\ L_i s(\theta_{1i}) + h_r \end{bmatrix}, \tag{10}$$

$$\lambda_i = \frac{i\pi}{3} + \frac{\pi}{6},$$

where θ_{1i} is the angle between the actuator and the fixed base. Since points C_1, C_2, and C_3 form an equilateral triangle in the top rotating platform, based on the theory in geometry, the relationship of $C_1C_2C_3$ can be determined by

$$\left|C_1C_2\right| = \left|C_1C_3\right| = \left|C_2C_3\right| = \sqrt{3}r_2. \tag{11}$$

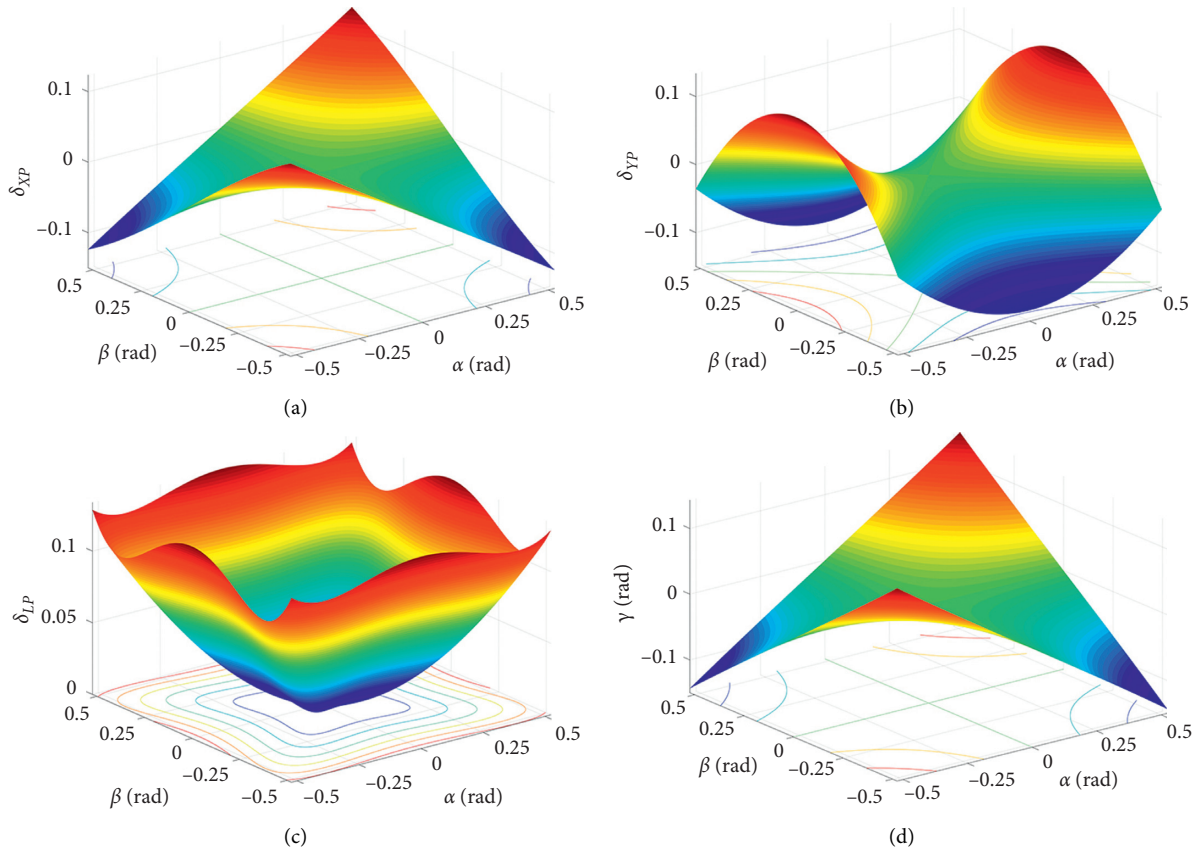


Figure 5: Influence factor.

Table 1: Geometric Parameters of the thrust-vectoring mechanism.

Symbol	Value (unit)
r_1	100 (mm)
r_2	60 (mm)
h	330 (mm)
h_r	10 (mm)

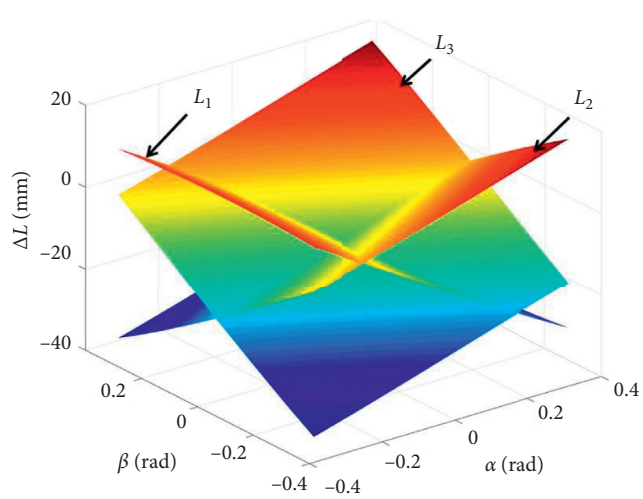


Figure 6: Length changes of linear actuators.

Hence, the position of points C_i in reference frame O-xyz can be obtained through (9)–(11).

Because P is the center of the circumcircle generated by three points $C_1C_2C_3$ in the top moving platform, the position of center point P in global reference frame O-xyz can be expressed as

$$\left|p - OC_i\right| = r_2, \tag{12}$$

where p of the center point P of the top rotating platform, which can be used to describe the position of point P with three components (P_x, P_y, P_z).

The position vector OC_i of points C_i is regarded as a known parameter when the length of linear actuator is given; thus, the three equations about the point of P can be established and calculated.

To further investigate the relationship between the lengths of linear actuators with the position and orientation of the top platform, different lengths of the linear actuator are used for the forward kinematic analysis. In this simulation, the linear actuator L_1, $L_2 \in [310\ 340]$ mm and $L_3 = 310$, 325, 340 mm. According to equations (9)–(12), the orientation and position vector of the top moving platform can be calculated directly, as shown in Figure 8.

3.3. Workspace Analysis of Thrust-Vectoring Mechanism. Due to the available space of AUV is limited, it is important to analyze the workspace of the thrust-vectoring

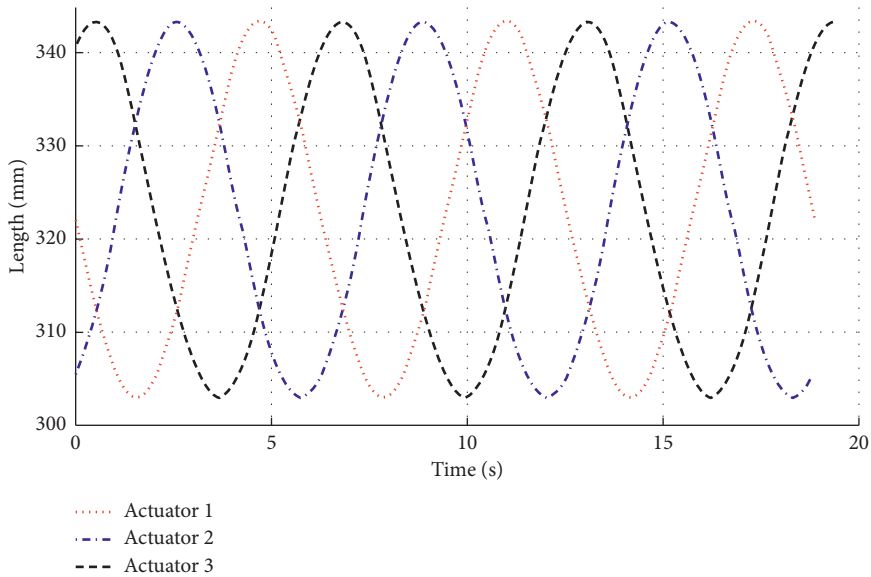


Figure 7: Length of linear actuators.

mechanism for optimizing structure design and improving performance. According to the kinematic analysis mentioned above, all the positions and orientations of the top rotating platform can be obtained by changing the lengths of actuators. Considering the motion characteristics of the thrust-vectoring mechanism and constraint on available space, the workspace analysis is mainly referring to study the tilt angle and angle change of the revolute joint and spherical joint of the thrust-vectoring mechanism in this paper.

In this section, θ_r and θ_t denote, respectively, the rotational angles of the revolute joint and spherical joint. The schematic diagram of revolute joint and spherical joint is presented, as shown in Figure 9.

The tilt angle θ_{ri} and angle change $\Delta\theta_{ri}$ of the revolute joint at point B_i can be defined by

$$\theta_{ri} = \arccos\frac{s_{ri}\cdot e_1}{|s_{ri}||e_1|},$$
$$\Delta\theta_{ri} = \theta_{ri} - \theta_{r-\text{avg}}, \tag{13}$$

where e_1 is the direction vector of the y-axis in global reference frame O-xyz and $\theta_{r-\text{avg}}$ is the initial angle generated by the linear actuator and the fixed base at the tilt angle $\alpha = \beta = 0$ rad.

The tilt angle θ_{ti} and angle change $\Delta\theta_{ti}$ of the spherical joint at point C_i can be expressed as

$$\theta_{ti} = \arccos\frac{s_{ti}\cdot t_1}{|s_{ti}||t_1|},$$
$$\Delta\theta_{ti} = \theta_{ti} - \theta_{t-\text{avg}}, \tag{14}$$

where t_1 is the direction vector of the k-axis in local reference frame P-ijk and $\theta_{t\text{-avg}}$ is the initial angle between the linear actuators with the rotating platform at the tilt angle $\alpha = \beta = 0$ rad.

Similarly, to study the relations between the tilt angles of revolute and spherical joints and the lengths of linear actuators, the tilt angles of revolute joint θ_{ri} and spherical joint θ_{ti} by forward kinematic analysis are performed with linear actuator length L_1, $L_2 \in [310\ \ 340]$ mm and $L_3 = 310$, 325, 340 mm. The simulation results are plotted in Figure 10.

The tilt angle of the three linear actuators and the top rotating platform has close relation, the platform moves according to α_s and β_s mentioned above. When the top rotating platform moves according to $\alpha = \alpha_s$ and $\beta = \beta_s$ ($\alpha_s = \pi/9 \cdot \sin(\pi \cdot (t/36)$ rad and $\beta_s = \pi/9 \cdot \cos(\pi \cdot t/36)$ rad), the tilt angle and angle change of linear actuators are plotted in Figure 11.

4. Dynamic Analysis of Thrust-Vectoring Mechanism

For improving the dynamic performance and control accuracy of the designed vectored thruster, it is greatly important to analyze the dynamics model. Since the thrust-vectoring mechanism is designed based on the 3RPS parallel manipulator, which includes three closed-loops kinematic chains, it is very complicated to perform the dynamic analysis of the thrust-vectoring mechanism.

According to the theoretical analysis and practical needs, the dynamic model of the thrust-vectoring mechanism based on the 3RPS parallel manipulator is established. The schematic diagram of the dynamic analysis model of the 3RPS parallel manipulator is represented graphically, as shown in Figure 12.

Referring to Figure 12, dynamic formulation of the force and moment balances on the linear actuator of the 3RPS parallel manipulator can be expressed as follows:

$$\begin{cases} F_{Bi} + F_{gi} - F_{bi} - F_{Ci} = m_t \dot{v}_{ti}, \\ M_{Bi} - M_{gi} - M_{Ci} + M_{bi} - m_t r_{ti} \times \dot{v}_{ti} = I_i \dot{\omega}_i + \omega_i \times I_i \omega_i, \end{cases} \tag{15}$$

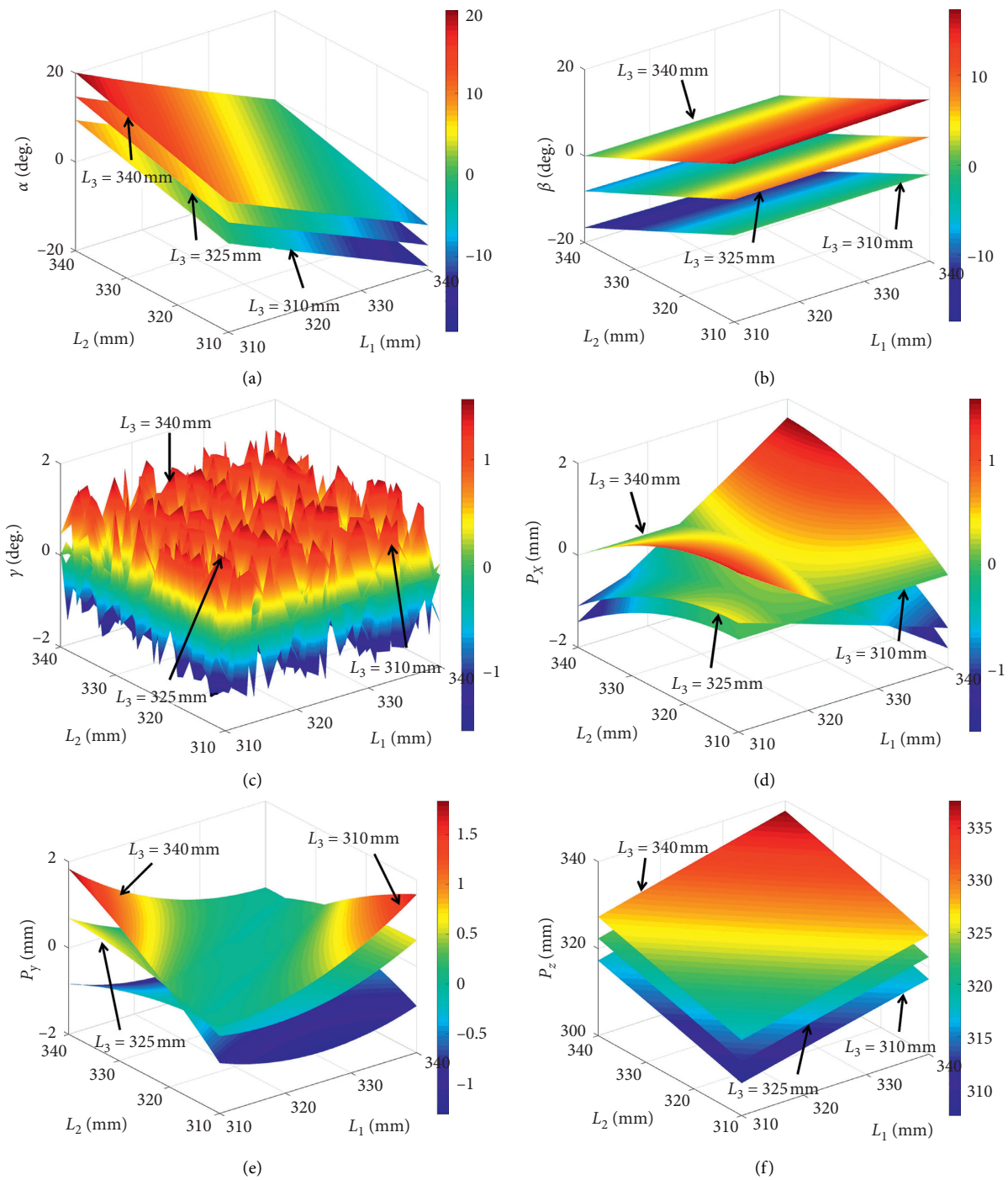


Figure 8: The position and orientation of the top rotating platform with certain actuator lengths.

where F_{Bi} and M_{Bi} denote force and moment applied at point B_i, and F_{Ci} and M_{Ci} denote force and moment at point C_i accordingly. F_{gi} and F_{bi} are the gravity and buoyancy of the linear actuator, M_{gi} and M_{bi} denote the moments generated by gravity and buoyancy of the linear actuator, respectively. It should be noted that the buoyancy of the linear actuator F_{bi} can be obtained by the diameters d_r and d_t of bottom section and the movable part and the length of bottom section l_r. m_t and I_i represent the mass of the translating component and the inertia moment of the linear actuator. $\dot{v}_{ti}$ represents the acceleration velocity of the linear actuator. r_{ti} denotes the vector from point B_i to the mass center of linear actuator, and $\dot{\omega}_i$ denote the angular velocity and acceleration of the linear actuator, respectively.

Due to acting by an external force and moment, it is necessary to carrying out dynamic analysis of the top moving platform for establishing overall dynamics model for the

Figure 9: Schematic diagram of revolute joint and spherical joint.

thrust-vectoring mechanism. According to the forces and moments distributions analysis of the proposed thrust-vectoring mechanism, the stress conditions of the top platform can be represented as Figure 13.

As shown in Figure 13, forces and moments are applied on the connection points C_i generated from many respects, such as the linear actuator, the top rotating platform, and duct propeller. Based on the definition of forces and moments at point C_i above, the dynamical equations of the top rotating platform can be expressed as

$$\begin{cases} \sum_{i=1}^{3} F_{Ci} + F_{g-P} - F_{b-P} + RF_e = m_P \dot{v}_P, \\ \sum_{i=1}^{3} (M_{Ci}) - l_{g-P} \times F_{g-P} + l_{b-P} \times F_{b-P} + RM_e = I_P \dot{\omega}_P + \omega_P \times (I_P \omega_P), \end{cases} \tag{16}$$

where v_p and $\dot{v}_P$ are the velocity and acceleration and ω_P and $\dot{\omega}_P$ denote the angular velocity and angular acceleration at the center of the top platform, respectively. F_{Ci} denotes actuating force from the linear actuator along the direction of actuator. Because the top platform is an axisymmetric structure, F_{g-P} and F_{b-P} are the gravity and buoyancy from the top platform, and l_{g-P} and l_{b-P} denote the distance from the center point P generated by connection points $C_1C_2C_3$ to the center of mass and buoyancy of the top rotating platform, respectively. F_e and M_e are the external force and moment mainly generated from the propeller and the duct in our paper. Referring to Figure 13, the external force F_e and external moment M_e can be defined as

$$\begin{cases} F_e = F_{\text{prop}} + F_{\text{duct}}, \\ M_e = M_{\text{duct}}, \end{cases} \tag{17}$$

where F_{prop} and F_{duct} denote the force generated by the propeller and the duct of the thrust-vectoring mechanism and M_e denotes the moment generated by the duct, respectively. The thrust vector F_{prop} is produced by the propeller and can be expressed as

$$F_{\text{prop}} = T_p \begin{bmatrix} -c\alpha c\beta \\ c\alpha s\beta \\ s\alpha \end{bmatrix}, \tag{18}$$

where T_p denotes the thrust produced by propeller and based on standard propeller theory [28, 31], $T_p = K_T \rho n^2{}_p D^4$. K_T, ρ, n_p, and D denote the thrust coefficient, the water density, the rotation speed of the propeller, and the propeller diameter, respectively.

In this AUV, the duct propeller has been widely adopted to protect from damage and improve the propulsive efficiency by being enclosed by a duct. To further investigate the dynamic model of the vectored thruster, it is clearly necessary to considerate the effect on lift and drag generated by the duct. The force generated by the duct applied to the platform can be expressed as

$$F_{LD} = \begin{bmatrix} 0 \\ -L \\ D \end{bmatrix}, \tag{19}$$

where L and D denote the lift and drag of the duct, which can be calculated by CFD software.

Because the duct rotates around the center of the duct in use, a transformation matrix R_d is established to convert the duct frame into the body frame, and the matrix R_d can be described as

$$R_d = \begin{bmatrix} c\alpha c\beta & -s\alpha & c\alpha s\beta \\ s\alpha c\beta & c\alpha & s\alpha s\beta \\ -s\beta & 0 & c\beta \end{bmatrix}, \tag{20}$$

where α and β represent the tilt angles of the ducted propeller.

Referring to equations (17)–(20), the force F_{duct} and moment M_{duct} generated by the duct that are applied on the platform of the thrust-vectoring mechanism can be calculated as

$$\begin{cases} F_{\text{duct}} = R_d F_{LD}, \\ M_{\text{duct}} = r_p \times F_{\text{duct}}. \end{cases} \tag{21}$$

Finally, based on the above analysis and according to Figures 12 and (15)–(21), the force balance along the leg direction can be expressed as

$$F_i = F_{Bi} \cdot s_i, \tag{22}$$

where F_i denotes the force produced by the linear actuator to complete the key components of drive function and s_i is the unit vector of the ith actuator.

In order to do a better research on the effect of motion on the vectored thruster, the numerical dynamic simulation on the thrust-vectoring mechanism has been developing. Some parameters used in the simulation, such as the dimension parameters of vectored thruster, are calculated through CAD software, and other parameters can be obtained by in [56]. All parameters of the thrust-vectoring mechanism are given in Table 2.

Based on the abovementioned theory analysis and parameters, the analysis formulations in Section 4 have been implemented in MATLAB.

When only considering the thrust produced by the duct propeller and the top platform moving according to $\alpha = \alpha_s$ and $\beta = 0$, the length changes and the driving forces of linear

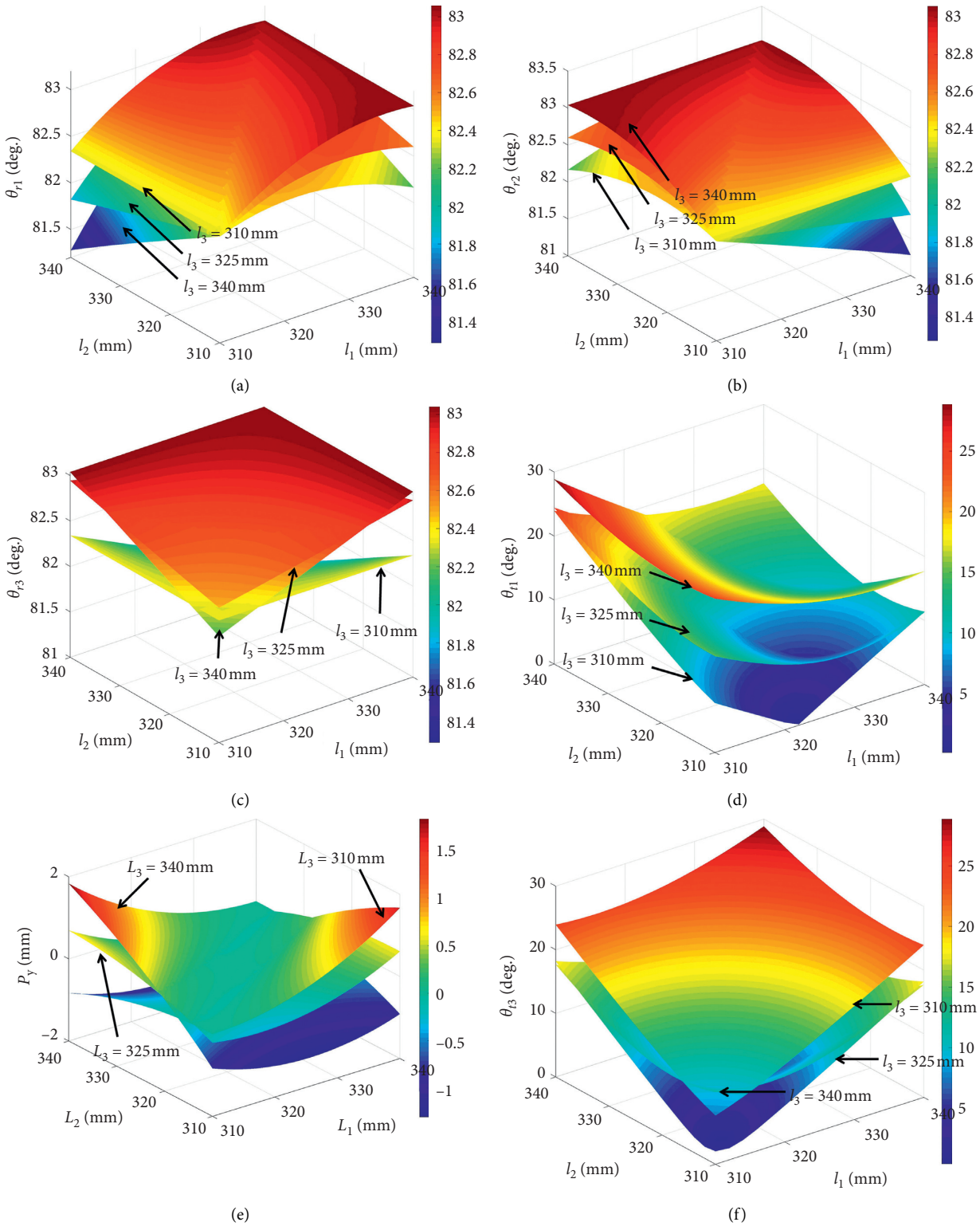


Figure 10: Tilt angle of revolute and spherical joint of the actuator by forward kinematic analysis.

actuator can be calculated, and the results are plotted in Figures 14 and 15.

In the dynamic analysis above, the gravity, buoyancy, drag, and torque of the vectored thruster is ignored. When the weights of the rotating platform and the three actuators are taken into account only, the driving forces of the linear actuator can be calculated and plotted in Figure 16.

In addition, the buoyancy of actuator depends on the length changes of actuators, and the buoyancy of linear actuator can be calculated by the length change. When the

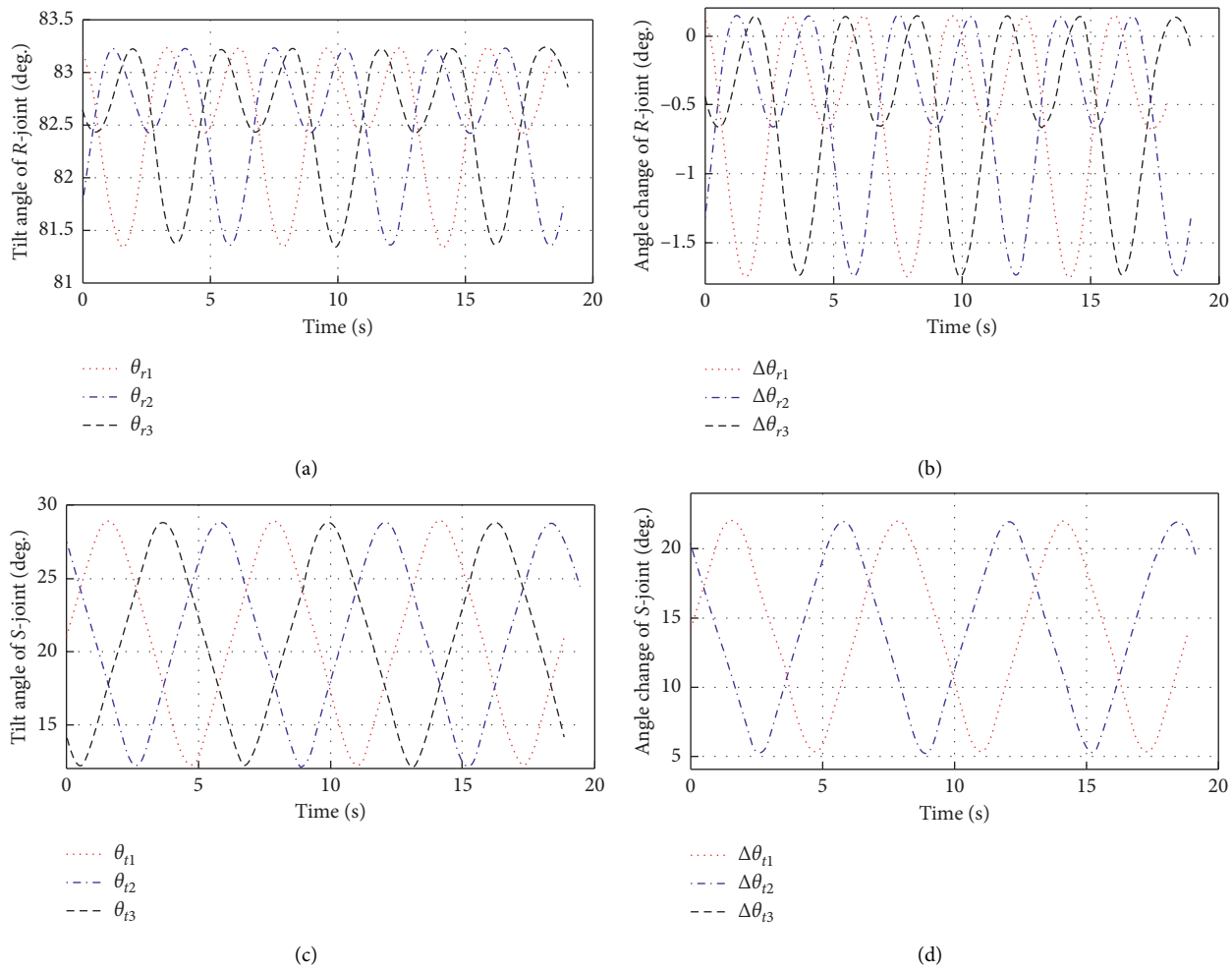


Figure 11: The tilt angle and angle change. (a) θ_{ri}. (b) $\Delta\theta_{ri}$. (c) θ_{ti}. (d) $\Delta\theta_{ti}$.

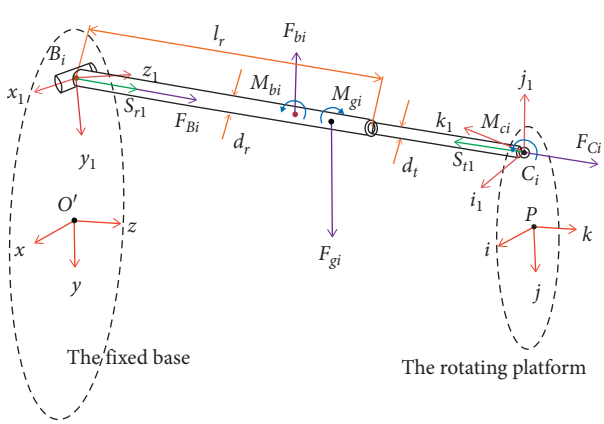


Figure 12: Schematic diagram of the dynamic model of the thrust-vectoring mechanism.

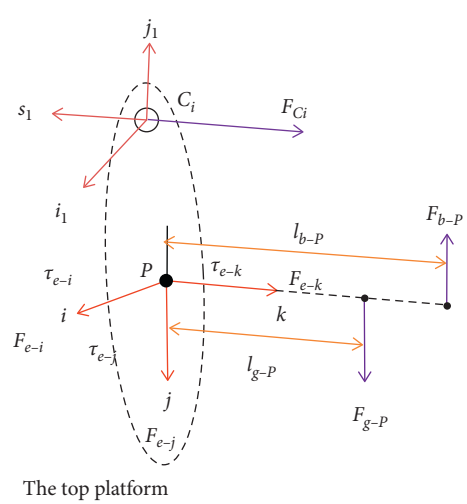


Figure 13: Schematic diagram of the dynamic analysis of the rotating platform.

top platform moves according to the designed trajectory, the buoyancy is created with the movement of the vectored thruster and the result of buoyancy is shown in Figure 17.

As we can see in Figure 17, the buoyant forces of actuators change with the length changes of actuators, but the magnitude of the change is relatively small. Hence, the buoyant forces of actuators can be approximately equal to 4.7 N in this next calculation.

Table 2: Parameters of the thrust-vectoring mechanism.

Symbol	Value (unit)
d_r	50 (mm)
d_t	10 (mm)
l_r	250 (mm)
l_{g-P}	35 (mm)
l_{b-p}	70 (mm)
F_{g-P}	100 (N)
F_{b-p}	30 (N)
F_{gi}	20 (N)
T_p	200 (N)
I_p	diag (3.2 3.2 1.54) (kg · m^2)
I_i	diag (24.5 24.5 0.1) (kg · m^2)
m_t	0.1 (kg)
P	60000
I	40000
D	15000

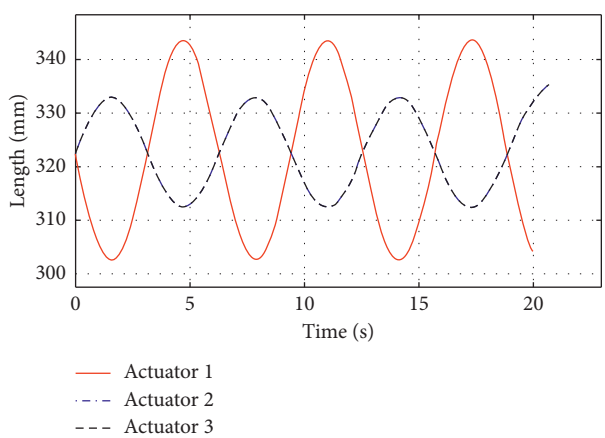


Figure 14: Length change of actuators at $\alpha = \alpha_s$ and $\beta = 0$.

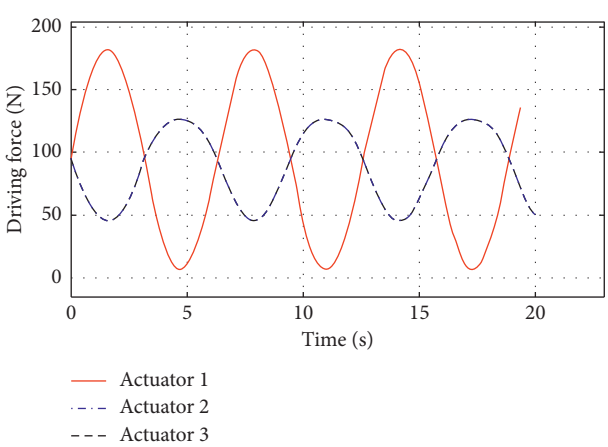


Figure 15: Driving forces of actuators at $\alpha = \alpha_s$ and $\beta = 0$.

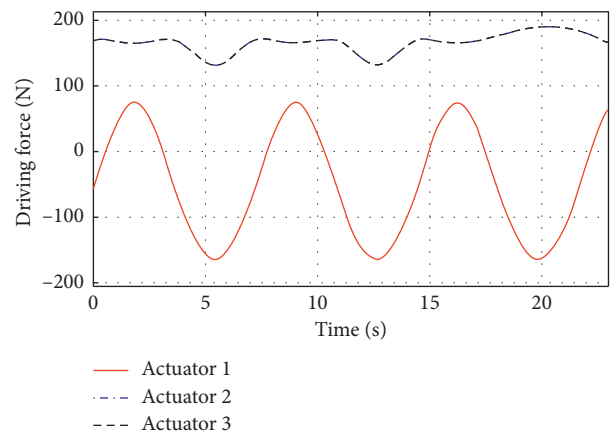


Figure 16: Driving forces of actuators at $\alpha = \alpha_s$ and $\beta = 0$.

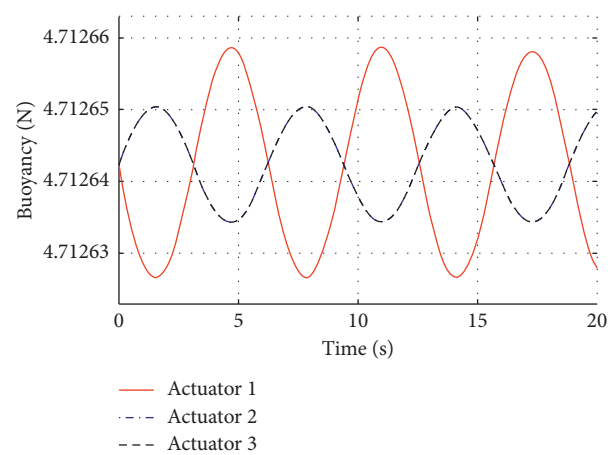


Figure 17: Buoyant forces of actuators at $\alpha = \alpha_s$ and $\beta = 0$.

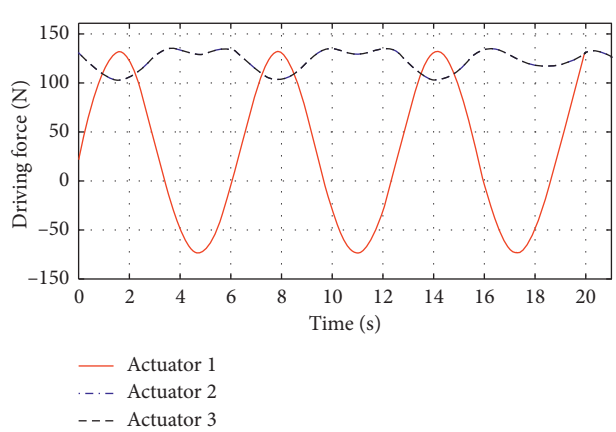


Figure 18: Driving forces of actuators at $\alpha = \alpha_s$ and $\beta = 0$.

Generally, the emphases of dynamic analysis of the parallel manipulator for research are mainly focused on the gravity, the external force, and torques. However, the buoyancy is also an important factor that affects the driving force of the actuators because the designed vectored thruster is used in underwater vehicles. In addition, the buoyancy of the other parts can be directly calculated by CAD software.

With considering the buoyant forces of actuators as shown in Figure 17 and buoyancy of other parts, the driving force of the ith actuator can be recalculated and plotted in Figure 18.

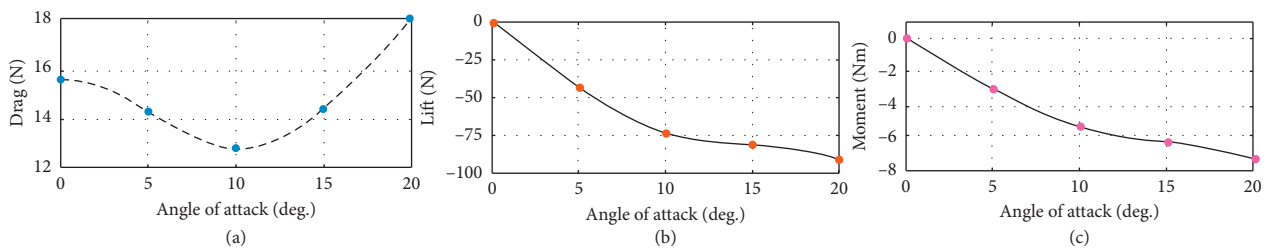


Figure 19: Drag (a), lift (b), and pitch moment (c) with different angles of attack.

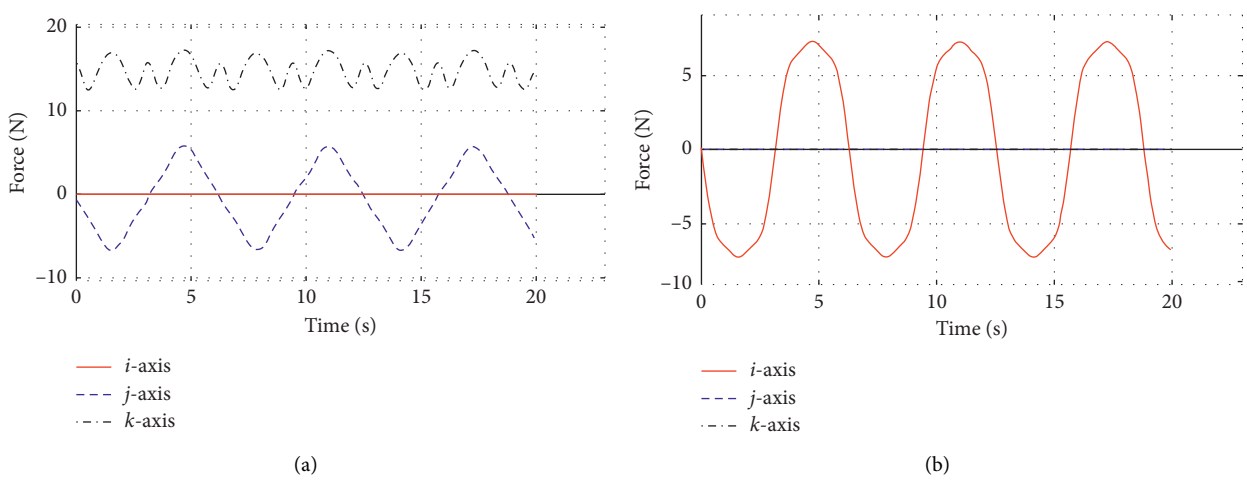


Figure 20: Force (a) and moment (b) generated by the duct at $\alpha = \alpha_s$ and $\beta = 0$.

Comparing the driving forces in Figures 15, 17, and 18, it is obvious that the gravity and buoyancy of the whole vectored thruster have an effect on the value of the driving force greatly.

In order to investigate the performance and characteristic of the vectored thruster more fully, some influential factors of the duct have been considered in the following simulations. Referring to (19)–(21), the parameters of the duct are important to calculate in the numerical dynamic simulation. In this paper, we use CFD simulation to get the dynamics parameters of the duct. Figure 19 shows the drag, lift, and moment of the duct with the angle of attack range between 0° and 20° and maintains the flow velocity at 2.5 m/s, respectively.

Based on the dynamics parameters of the duct, as shown in Figure 19, the force and moment generated by the duct are calculated and presented in Figure 20.

The calculation results show that the influential factors of the duct play a very important role in calculating the driving forces of actuators; hence, the influential factors of the duct should be considered in a calculation schedule.

To improve the performance of the vehicle motion control, the tilt angles α and β play a crucial role in the vectored thruster AUVS. With considering the usage of the vectored thruster, a control scheme using the PID method is developed for the designed thrust-vectoring mechanism, as shown in Figure 21. PID algorithm is the most widely used control methods in all kinds of application fields for its effectiveness and practicability. Based on the control scheme shown in Figure 21, a control model of the proposed thrust-vectoring mechanism is developed by Matlab and Simscape Mutibody, as shown in Figures 22 and 23.

To further investigate the performance of proposed control model, related numerical simulations are carried out with top platform moves according to $= \alpha_s$ and $\beta = 0$. The related parameters are given in Table 2, the length responses of linear actuators can be obtained, as shown in Figure 24.

The simulation results from Figures 14 and 24 show that the designed PID controller for the thrust-vectoring mechanism is fast, effective, and able to achieve the expected goal commendably. Based on the designed controller and considering the influence of factors as analyzed above, which includes extra forces and moments from duct and buoyant forces of the vectored thruster, the driving force of linear actuators is presented in Figure 25.

Comparing the driving force in Figures 18 and 25, the driving forces of actuators have some similarities in change trend at the same time, but the maximum and minimum of driving forces are distinctly different. Hence, it is concluded that the lift, drag, and torque of the duct propeller are important influential factors to the dynamic model of the vectored thruster. Accordingly, for the purposes of optimizing the structure and decreasing the dimension, it is of significance to choose the appropriate linear actuator for driving the proposed thrust-vectoring mechanism by studying the dynamic analysis with considering all kinds of influence factor.

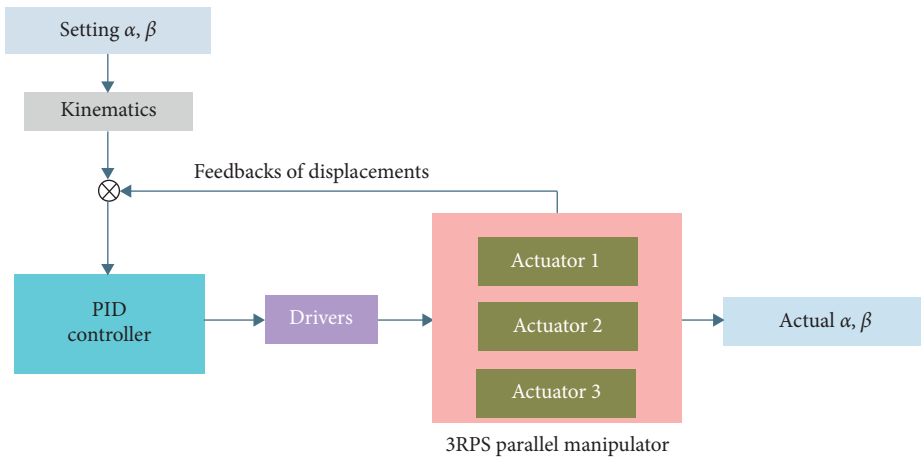


Figure 21: Control scheme of the thrust vector control system.

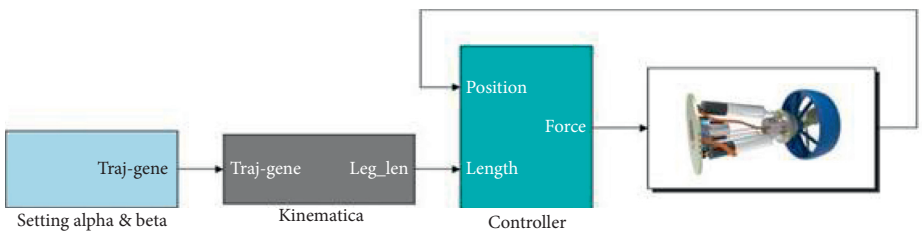


Figure 22: Control scheme of the thrust-vectoring mechanism.

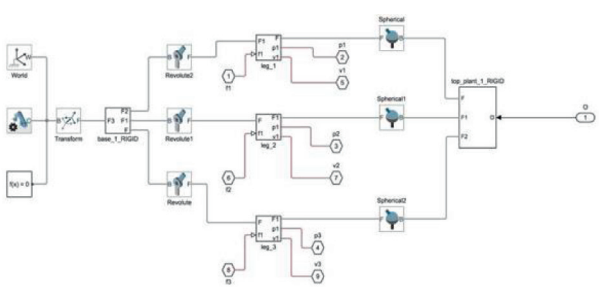


Figure 23: Model of the the thrust-vectoring mechanism with Simscape Mutibody.

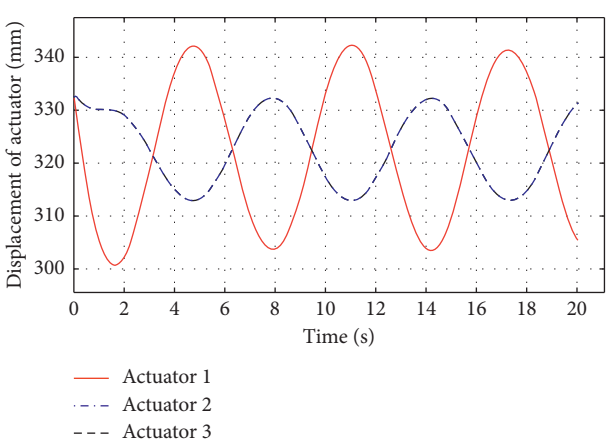


Figure 24: The length response of the linear actuator.

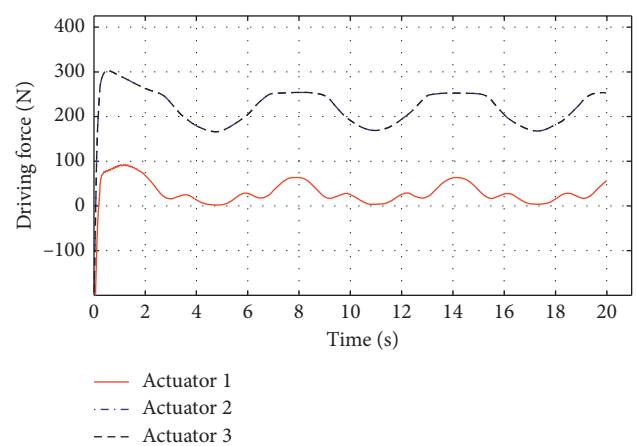


Figure 25: Driving forces of linear actuator.

5. Conclusion

In this paper, a design scheme for the vectored thruster based on a 3RPS parallel manipulator is proposed to solve the effect of the control surface weakening problems. Parallel manipulators have several advantages over the mechanical structure and are suitable for various application fields, such as compact and reasonable structure, fast response, and high positioning precision. Because of the merits of itself, this type mechanical structure is used to design the thrust-vectoring mechanism considering the limited space and hash environment. Additionally, a duct propeller is adopted

as power source of the proposed vectored thruster, which is installed on the top rotating platform with a main motor as a whole structure, thereby this design ensures compact structure, reliable motion, and high propulsive efficiency. Because the control force is provided by the component force of thrust rather than common rudders, the proposed vectored thruster AUVs have the abilities to complete all kinds of certain tasks and operations at a low forward speed.

In order to make sure the designed vectored thruster can run efficiently and stably, studying and developing the control system is fundamental to implement the design function of the vectored thruster. Owing to the importance of control system, related theoretical research about kinematics and dynamics of the thrust-vectoring mechanism is carried out to establish the motion model. In the kinematic analysis, the inverse and forward kinematics of the thrust-vectoring mechanism is presented, and the numerical simulations are accomplished to prove the feasibility and effectiveness of this design. In the section of workspace analysis, the study of the tilt angles of revolute and spherical joints is also carried out to make sure the motion platform can implement its designed function in limited motion space. In order to further verify feasibility of the thrust-vectoring mechanism, based on the considerations of various affecting factors, a dynamics model of the designed thrust-vectoring mechanism is established according to theoretical analysis, and the driving forces of the linear actuator are presented though a series of numerical simulations. In addition, a control scheme based on PID algorithm is proposed for the thrust vector control system on the existing work basis, and a control model is established using Simsacpe Mutibody, and the simulation results proved the feasibility of the proposed control scheme, which can effectively realize the goal of controlling the thrust-vectoring mechanism.

According to the above, the designed vectored thruster is able to provide the vectored thrust effectively and efficiently, and the AUVs equipped with vectored thruster are able to complete a variety of the complex tasks at a comparably low forward speed.

In the future research, a series of numerical simulations and theoretical study are carried out to investigate hydrodynamic performance of this vectored thruster AUV. On this basis, a prototype of this designed vectored thruster will be developed and experimental test will be carried out to verify the principles design. Moreover, the corresponding control system of the vectored thruster as a part of the AUV will be developed and tested in pools or open water to check its performance.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 51879220).

References

[1] V. Upadhyay, S. Gupta, A. C. Dubey et al., "Design and motion control of autonomous underwater vehicle," in *Proceedings of the 2015 IEEE Underwater Technology (UT)*, Chennai, India, February 2015.

[2] J. Mccolgan and E. Mcgookin, "Coordination of multiple biomimetic autonomous underwater vehicles using strategies based on the schooling behaviour of fish," *Robotics*, vol. 5, no. 1, p. 2, 2016.

[3] M. Furlong, D. Paxton, P. Stevenson, M. Pebody, S. D. Mcphail, and J. Perrett, "Autosub long range: a long range deep diving AUV for ocean monitoring," in *Proceedings of the 2012 IEEE/OES Autonomous Underwater Vehicles (AUV)*, Southampton, UK, September 2012.

[4] B.-H. Jun, J.-Y. Park, F.-Y. Lee et al., "Development of the AUV "ISiMI" and a free running test in an ocean engineering basin," *Ocean Engineering*, vol. 36, no. 1, pp. 2–14, 2009.

[5] Y. Zhang, Y. Li, Y. Sun, J. Zeng, and L. Wan, "Design and simulation of x-rudder AUV'S motion control," *Ocean Engineering*, vol. 137, pp. 204–214, 2017.

[6] S. K. Jain, S. Mohammad, S. Bora, and M. Singh, "A review paper on: autonomous underwater vehicle," *International Journal of Scientific & Engineering Research*, vol. 6, no. 2, 2015.

[7] J. Elvander and G. Hawkes, "ROVs and AUVs in support of marine renewable technologies," in *Proceedings of the 2012 Oceans*, Hampton Roads, VA, USA, October 2012.

[8] Y. Allard and E. Shahbazian, "Unmanned underwater vehicle (UUV) information study," Technical report, OODA Technologies Inc, Montreal, Canada, 2014.

[9] Y. Wang, B. Jiang, Z.-G. Wu, S. Xie, and Y. Peng, "Adaptive sliding mode fault-tolerant fuzzy tracking control with application to unmanned marine vehicles," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–10, 2020.

[10] T. I. Fossen and O.-E. Fjellstad, "Robust adaptive control of underwater vehicles: a comparative study," *IFAC Proceedings Volumes*, vol. 28, no. 2, pp. 66–74, 1995.

[11] Y. Wang, C. K. Ahn, H. Yan, and S. Xie, "Fuzzy control and filtering for nonlinear singularly perturbed Markov jump systems," *IEEE Transactions on Cybernetics*, pp. 1–12, 2020.

[12] Y. Wang, H. R. Karimi, H.-K. Lam, and H. Yan, "Fuzzy output tracking control and filtering for nonlinear discrete-time descriptor systems under unreliable communication links," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2369–2379, 2019.

[13] Y. Wang, X. Xie, M. Chadli, S. Xie, and Y. Peng, "Sliding mode control of fuzzy singularly perturbed descriptor systems," *IEEE Transactions on Fuzzy Systems*, p. 1, 2020.

[14] T. Ura, T. Obara, S. Takagawa, and T. Gamo, "Exploration of Teisi knoll by autonomous underwater vehicle R-one robot," in *Proceedings of the MTS/IEEE Oceans 2001. An Ocean Odyssey*, Honolulu, HI, USA, November 2001.

[15] P. B. Sujit, J. B. De Sousa, and F. L. Pereira, "UAV and AUVS coordination for ocean exploration," in *Proceedings of the OCEANS 2009-EUROPE*, Bremen, Germany, May 2009.

[16] K. Weitemeyer and S. Constable, "Mapping shallow geology and gas hydrate with marine CSEM surveys," *First Break*, vol. 28, no. 6, 2010.

[17] P. Wessel and M. T. Chandler, "The spatial and temporal distribution of marine geophysical surveys," *Acta Geophysica*, vol. 59, no. 1, pp. 55–71, 2011.

[18] M. Grasmueck, G. P. Eberli, D. A. Viggiano, T. Correa, G. Rathwell, and J. Luo, "Autonomous underwater vehicle (AUV) mapping reveals coral mound distribution, morphology, and oceanography in deep water of the straits of Florida," *Geophysical Research Letters*, vol. 33, no. 23, 2006.

[19] S. A. T. Randeni, A. Forrest, R. Cossu, Z. Leong, D. Ranmuthugala, and V. Schmidt, "Parameter identification of a nonlinear model: replicating the motion response of an autonomous underwater vehicle for dynamic environments," *Nonlinear Dynamics*, vol. 91, no. 2, pp. 1229–1247, 2018.

[20] M. D. Henschel, R. Olsen, P. Hoyt, and P. W. Vachon, "The ocean monitoring workstation: experience gained with radarsat," in *Proceedings of the 1997 Geomatics in the ERA of RADARSAT (GER'97)*, Ottawa, Canada, 1997.

[21] D. C. Webb, P. J. Simonetti, and C. P. Jones, "SLOCUM: an underwater glider propelled by environmental energy," *IEEE Journal of Oceanic Engineering*, vol. 26, no. 4, pp. 447–452, 2001.

[22] L. A. Henkel, H. Nevins, M. Martin, S. Sugarman, J. T. Harvey, and M. H. Ziccardi, "Chronic oiling of marine birds in California by natural petroleum seeps, shipwrecks, and other sources," *Marine Pollution Bulletin*, vol. 79, no. 1-2, pp. 155–163, 2014.

[23] N. Gracias, P. Ridao, R. Garcia et al., "Mapping the Moon: using a lightweight AUV to survey the site of the 17th century ship La Lune," in *Proceedings of the OCEANS-Bergen 2013*, Bergen, Norway, June 2013.

[24] K. Kornei, "Seafloor data from lost airliner search are publicly released," *EOS*, vol. 98, 2017.

[25] M. Purcell, D. Gallo, G. Packard et al., "Use of REMUS 6000 AUVs in the search for the air France flight 447," in *Proceedings of the OCEANS 2011*, Waikoloa, HI, USA, September 2011.

[26] M. H. Khodayari and S. Balochian, "Modeling and control of autonomous underwater vehicle (AUV) in heading and depth attitude via self-adaptive fuzzy PID controller," *Journal of Marine Science and Technology*, vol. 20, no. 3, pp. 559–578, 2015.

[27] D. R. Blidberg, "The development of autonomous underwater vehicles (AUV); a brief summary," *IEEE ICRA*, vol. 4, p. 1, 2001.

[28] K. Tanakitkorn, P. A. Wilson, S. R. Turnock, and A. B. Phillips, "Depth control for an over-actuated, hover-capable autonomous underwater vehicle with experimental verification," *Mechatronics*, vol. 41, pp. 67–81, 2017.

[29] B. Xin, L. Xiaohui, S. Zhaocun, and Z. Yuquan, "A vectored water jet propulsion method for autonomous underwater vehicles," *Ocean Engineering*, vol. 74, pp. 133–140, 2013.

[30] T. Liu, Y. Hu, H. Xu, Z. Zhang, and H. Li, "Investigation of the vectored thruster AUVS based on 3SPS-s parallel manipulator," *Applied Ocean Research*, vol. 85, pp. 151–161, 2019.

[31] K. Tanakitkorn, P. A. Wilson, S. R. Turnock, and A. B. Phillips, "Sliding mode heading control of an overactuated, hover-capable autonomous underwater vehicle with experimental verification," *Journal of Field Robotics*, vol. 35, no. 3, pp. 396–415, 2018.

[32] X. Lin and S. Guo, "Development of a spherical underwater robot equipped with multiple vectored water-jet-based thrusters," *Journal of Intelligent & Robotic Systems*, vol. 67, no. 3-4, pp. 307–321, 2012.

[33] I. Carlucho, M. De Paula, S. Wang, Y. Petillot, and G. G. Acosta, "Adaptive low-level control of autonomous underwater vehicles using deep reinforcement learning," *Robotics and Autonomous Systems*, vol. 107, pp. 71–86, 2018.

[34] T. Wang and B.-w. Song, "Control of dynamic position system for AUV with multiple thrusters," *Acta Armamentarii*, vol. 5, 2006.

[35] T. Liu, Y. Hu, H. Xu, Q. Wang, and W. Du, "A novel vectored thruster based on 3-RPS parallel manipulator for autonomous underwater vehicles," *Mechanism and Machine Theory*, vol. 133, pp. 646–672, 2019.

[36] R. Panish, "Dynamic control capabilities and developments of the Bluefin robotics AUV fleet," in *Proceedings of the 16th International Symposium on Unmanned Untethered Submersible Technology (UUST)*, Durham, NH, USA, 2009.

[37] L. E. Bird, A. Sherman, and J. Ryan, "Development of an active, large volume, discrete seawater sampler for autonomous underwater vehicles," in *Proceedings of the OCEANS 2007*, Vancouver, Canada, September 2007.

[38] Y. Zhang, R. S. McEwen, J. P. Ryan et al., "A peak-capture algorithm used on an autonomous underwater vehicle in the 2010 Gulf of Mexico oil spill response scientific survey," *Journal of Field Robotics*, vol. 28, no. 4, pp. 484–496, 2011.

[39] H. Cheng, Y.-K. Yiu, and Z. Li, "Dynamics and control of redundantly actuated parallel manipulators," *IEEE/ASME Transactions on Mechatronics*, vol. 8, no. 4, pp. 483–491, 2003.

[40] D. Liang, Y. Song, and T. Sun, "Nonlinear dynamic modeling and performance analysis of a redundantly actuated parallel manipulator with multiple actuation modes based on FMD theory," *Nonlinear Dynamics*, vol. 89, no. 1, pp. 391–428, 2017.

[41] A. G. Ruiz, J. C. Santos, J. Croes, W. Desmet, and M. M. da Silva, "On redundancy resolution and energy consumption of kinematically redundant planar parallel manipulators," *Robotica*, vol. 36, no. 6, pp. 809–821, 2018.

[42] S. J. Yan, S. K. Ong, and A. Y. C. Nee, "Stiffness analysis of parallelogram-type parallel manipulators using a strain energy method," *Robotics and Computer-Integrated Manufacturing*, vol. 37, pp. 13–22, 2016.

[43] E. Cavallo, R. C. Michelini, and V. F. Filaretov, "Conceptual design of an AUV equipped with a three degrees of freedom vectored thruster," *Journal of Intelligent and Robotic Systems*, vol. 39, no. 4, pp. 365–391, 2004.

[44] L. Nurahmi, J. Schadlbauer, S. Caro, M. Husty, and P. Wenger, "Kinematic analysis of the 3-RPS cube parallel manipulator," *Journal of Mechanisms and Robotics*, vol. 7, no. 1, Article ID 011008, 2015.

[45] A. Nayak, T. Stigger, M. L. Husty, P. Wenger, and S. Caro, "Operation mode analysis of 3-RPS parallel manipulators based on their design parameters," *Computer Aided Geometric Design*, vol. 63, pp. 122–134, 2018.

[46] J. Gallardo, H. Orozco, and J. M. Rico, "Kinematics of 3-RPS parallel manipulators by means of screw theory," *The International Journal of Advanced Manufacturing Technology*, vol. 36, no. 5-6, pp. 598–605, 2008.

[47] N. M. Rao and K. M. Rao, "Multi-position dimensional synthesis of a spatial 3-RPS parallel manipulator," *Journal of Mechanical Design*, vol. 128, no. 4, pp. 815–819, 2006.

[48] H. S. Kim and L.-W. Tsai, "Kinematic synthesis of spatial 3-RPS parallel manipulators," in *Proceedings of the ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, pp. 873–880, 2002.

[49] A. Nayak, L. Nurahmi, P. Wenger, and S. Caro, "Comparison of 3-RPS and 3-SPR parallel manipulators based on their

maximum inscribed singularity-free circle," in *New Trends in Mechanism and Machine Science*, pp. 121–130, Springer, Berlin, Germany, 2017.

[50] Q. Li, J. n. Xiang, X. Chai, and C. Wu, "Singularity analysis of a 3-RPS parallel manipulator using geometric algebra," *Chinese Journal of Mechanical Engineering*, vol. 28, no. 6, pp. 1204–1212, 2015.

[51] D. Thompson, D. Caress, H. Thomas, and D. Conlin, "MBARI mapping AUV operations in the gulf of California 2015," in *Proceedings of the OCEANS 2015-MTS/IEEE Washington*, Washington, DC, USA, October 2015.

[52] A. Bhattacharyya and S. Steen, "Influence of ducted propeller on seakeeping in waves," *Ocean Engineering*, vol. 91, pp. 243–251, 2014.

[53] M. Díaz-Rodríguez, J. A. Carretero, and R. Bautista-Quintero, "Solving the dynamic equations of a 3-PRS parallel manipulator for efficient model-based designs," *Mechanical Sciences*, vol. 7, no. 1, p. 9, 2016.

[54] J. A. Carretero, R. P. Podhorodeski, M. A. Nahon, and C. M. Gosselin, "Kinematic analysis and optimization of a new three degree-of-freedom spatial parallel manipulator," *Journal of Mechanical Design*, vol. 122, no. 1, pp. 17–24, 2000.

[55] J. Carretero, M. Nahon, and R. Podhorodeski, "Workspace analysis and optimization of a novel 3-DOF parallel manipulator," *International Journal of Robotics and Automation*, vol. 15, no. 4, pp. 178–188, 2000.

[56] R. McEwen and K. Streitlien, "Modeling and control of a variable-length AUV," in *Proceedings of the 12th International Symposium on Unmanned Untethered Submersible Technology*, Durham, NJ, USA, 2001.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 7141237, 14 pages
https://doi.org/10.1155/2020/7141237

Research Article

Prediction Model and Experimental Study on Braking Distance under Emergency Braking with Heavy Load of Escalator

Zhongxing Li,[1] Haixia Ma,[2] Peng Xu,[3] Qifeng Peng,[1] Guojian Huang,[1] and Yingjie Liu[1]

[1]*Guangzhou Academy of Special Equipment Inspection & Testing, Guangzhou 510180, China*
[2]*Guangzhou College of South China University of Technology, Guangzhou 510800, China*
[3]*MOE Key Laboratory of Disaster Forecast and Control in Engineering, School of Mechanics and Construction Engineering, Jinan University, Guangzhou 510632, China*

Correspondence should be addressed to Haixia Ma; mahx@gcu.edu.cn and Peng Xu; 325802168@qq.com

Received 16 July 2020; Revised 8 August 2020; Accepted 17 August 2020; Published 2 September 2020

Academic Editor: Guoqiang Wang

In order to study the relationship between the braking distance and the load of escalator and realize the prediction of the rated load braking distance with a little load, the method of combining theoretical analysis and experimental research is used. First, the dynamic characteristics of the escalator during emergency braking are analyzed, and the prediction model of the braking distance of the escalator under different loads is derived based on the law of conservation of energy. Furthermore, the influence coefficients under different loads were determined through experimental studies, the model was revised, and the concept of equivalent no-load kinetic energy (ENKE) was proposed. The research shows that the braking distance of the escalator increases nonlinearly with the increase in load. When the no-load braking distance and the 25% rated load braking distance change greatly, the braking distance increases faster as the load increases; the escalators with large brake force have a small ENKE and are easy to stop. Otherwise, it is difficult to stop. The test results show that the comparison between the predicted value of the prediction model and the measured value has a maximum error of 2.7%, and the maximum error at rated load is only 2.0%, which fully meets the needs of engineering measurement. And the prediction method reduces test costs, enhances test security, and improves test coverage.

1. Introduction

As an important means of transportation in modern buildings, escalator greatly saves physical strength and time and improves traffic efficiency. As a kind of elevator, the escalator is used to transport passengers up or down in an oblique way. It has cyclic steps which are characterized by continuous operation. Compared with the vertical elevator, it has a greater transport capacity and is widely used in airports, shopping malls, stations, and other places with large passenger flow. In recent years, with the rapid development of economy, the number of elevators in China has increased rapidly, especially in the coastal areas. According to the market supervision administration of Guangdong Province website [1], by the end of 2019, the number of special equipment in Guangdong Province was 1.6356 million, 7.33% more than in 2018, 40800 boilers, 399700 pressure vessels, 853200 sets of elevator, 193500 sets of hoisting machinery, 146200 special motor vehicles, 24 passenger ropeways, and large-scale amusement equipment 2160 units (sets), as shown in Figure 1.

Due to the large amount and high risk, accidents of special equipment occur from time to time. In 2019, there were 15 special equipment accidents in Guangdong Province, including 9 elevator accidents, accounting for 60%. Figure 2 shows the accidents of special equipment in Guangdong Province in 2019. It can be seen that the elevator is not only a large number but also a special equipment with a high incidence of accidents.

Guangdong Provincial government attaches great importance to the safety of special equipment. With the contribution of technical personnel to safety technology, the

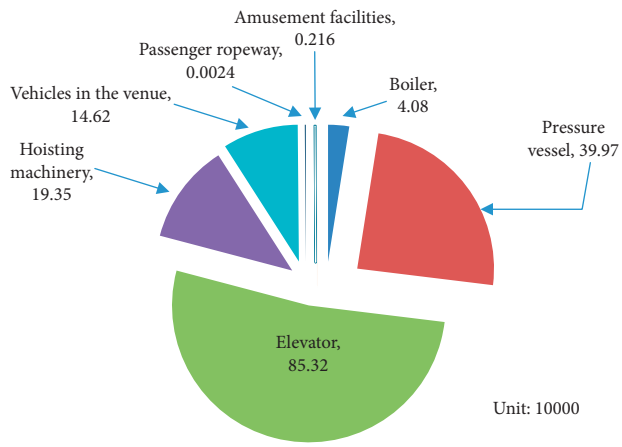
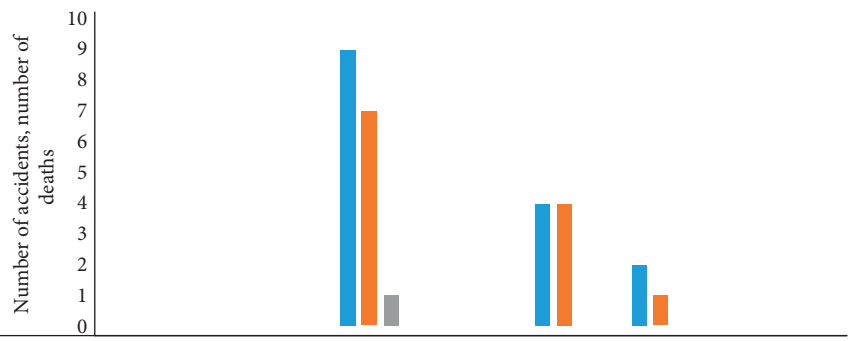


Figure 1: Proportion of special equipment in Guangdong Province in 2019.



Special equipment	Boiler	Pressure vessel	Elevator	Hoisting machinery	Vehicles in the venue	Amusement facilities	Passenger ropeway
Number of accidents	0	0	9	0	4	2	0
Number of deaths	0	0	7	0	4	1	0
Seriously injured people	0	0	1	0	0	0	0

Number of accidents
Number of deaths
Seriously injured people

Figure 2: Statistics of special equipment accidents in Guangdong Province in 2019.

deaths number of 10,000 special equipments in Guangdong Province has been decreasing year by year. However, due to the large base, the safety situation is still grim. Figure 3 shows the trend of accidents per 10,000 devices special equipment in Guangdong Province.

As a special equipment, the safety and reliability of escalators have always been the focus of managers and technicians [2–6]. Most escalator accidents occur in crowded public places, causing great personal injury and social impact [7–11]. Therefore, how to prevent accidents and reduce the loss and social impact brought by accidents has been the research direction of scholars in the industry [10, 12–14].

As an important part of escalator safety protection, the brake is used to quickly stop the escalator in the event of an emergency and protect the personal safety to the greatest extent. The principle of the brake is shown in Figure 4. The quality of the brake's performance directly affects the safety level of escalators [15–17]. Therefore, in order to improve the safety of escalator brake, domestic and foreign scholars have done a lot of research [18–20].

As an important indicator of brake performance, braking distance is a necessary test item for inspection and maintenance [21, 22]. GB16899-2011 5.4.2.1 has clearly stipulated the range of braking distance under different rated speeds and takes the detection items of braking distance into one of the items of escalator safety detection [23]. Table 1 shows the standard values of the braking distances of escalators at different nominal speeds.

In Table 1, the minimum braking distance is the limit value with no load, while the maximum braking distance is the limit value with rated load. The traditional rated load test

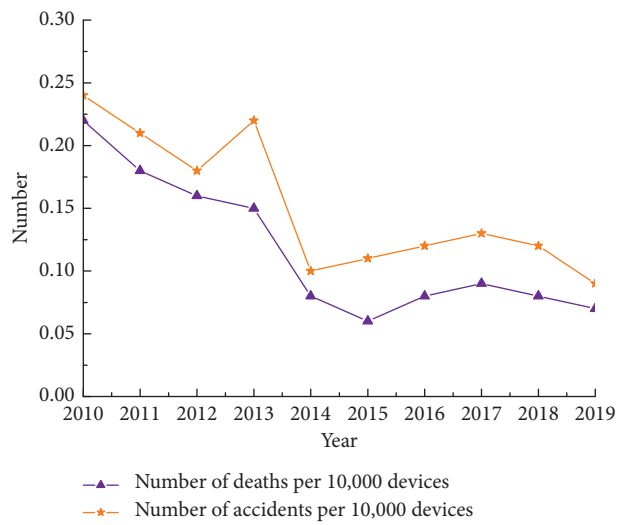


Figure 3: The trend of accidents of special equipment in Guangdong Province.

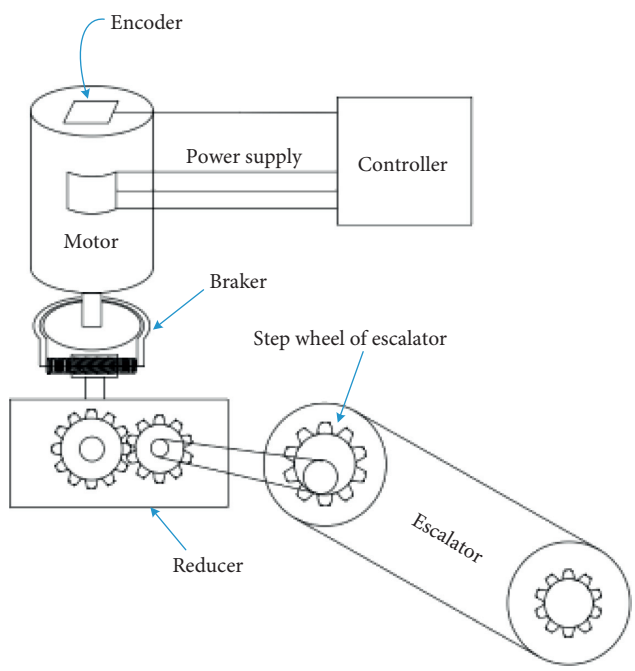


Figure 4: Structure principle of brake.

Table 1: Escalator braking distance range at different nominal speeds.

Nominal speeds V_0 (m/s)	Minimum braking distance S_{min} (m)	Maximum braking distance S_{max} (m)
0.5	0.2	1.0[a]
0.65	0.3	1.3[a]
0.75	0.4	1.5[a]

[a]Excluding endpoint values.

of the braking distance of escalators is mainly carried out by loading. When the escalator comes to an emergency stop under the rated load, the braking distance is measured to judge the braking performance of the brake. Rated load is required during the experiment. Due to the need to reserve the braking distance, the load will generally be concentrated on the upper part of the escalator, which can cause excessive load concentration. There are two problems in the load test: first, it is difficult to carry the weight that is required by the test, and the cost of handling is high; second, the brake

performance of the escalator with a long service life may not meet the requirements of the rated load test. And the escalator cannot be stopped reliably under rated load, which is a great potential safety hazard. The loading test situation is shown in Figure 5(a). Especially, in some shopping malls, there is a human activity space under the escalator, as shown in Figure 5(b). Once the brake is unable to stop the escalator reliably at rated load, in some cases, the escalator is damaged, and in other cases, personal injury or death will be caused.

Consider the high cost and risk of rated load test of braking distance. At present, for the annual inspection and maintenance of escalators, there is no rated load test, only the measurement of no-load braking distance. However, the actual situation is that escalators in China often operate under heavy load, as shown in Figure 5(c). This brings great safety hazard to passengers. In order to reduce the cost of rated load test and reduce the test risk, it is particularly important to study the equivalent test method under the condition of light load. By establishing the relationship model between the braking distance and load, the braking distance of escalator under heavy load will be predicted.

Some scholars have studied the relationship between the test of the braking distance and safety. Wang and Lu proposed a method for calculating the braking capacity of escalators and moving walkways. The braking torque is selected through braking deceleration and then proofread stop distance [24]. Based on the design requirements of the escalator braking system, Pan studied the influencing factors of the braking distance and put forward proposed improvements to the inspection requirements of the braking distance [25]. Hu established a new method for calculating the mathematical model of the escalator braking distance and used this method to calculate some structural parameters of the escalator [26]. Liu analyzed the calculation of the braking distance of the escalator and established the relevant calculation formula [27]. Park and Gschwendtner proposed an efficient multibody dynamics simulation modeling approach. The approach also covers a comprehensive simulation modeling of drive machine with gearbox, main drive chain band, operational brake system, and auxiliary brake system to evaluate the escalator brake performance at the system level [21]. The work of the researchers provided an important reference for the study of the braking distance of escalators. However, most of the research on the braking distance of escalators is carried out from the aspects of checking and influencing factors, and there are few studies on how to predict the braking distance under various loads without loading or adding light load. In view of this, this paper studies the law of movement in the process of escalator operation, analyzes the energy change during the escalator braking process, analyzes and summarizes a large number of test data, and puts forward a method for braking distance prediction.

In this paper, the braking distance prediction model of the escalator under heavy load is obtained by light load. The method can reduce the test cost and risk, increase the coverage of the braking distance test of rated load, and improve the safety of the escalator. The structure of the paper is as follows: Section 2 is the theoretical analysis, starting from the general equation of mechanical motion, analyzes the law of the movement of the escalator, and constructs the equivalent dynamic model of the escalator system. And the relationship between the braking distance and the load is formed. Section 3 establishes the escalator braking distance prediction model. Through the analysis of the law of energy change in the braking process and according to the conservation of energy, a simpler relationship between the braking distance and the load is deduced and tested. Section 4 is the improved prediction model of the braking distance of the escalator. The prediction model derived in the Section 3 is improved, and the calculation method of the influence coefficient is given. Section 5 is the calculation results of the improved model. Three escalators with different parameters are tested, and compared with the predicted data, the results are basically consistent to meet the needs of engineering measurement. In the fifth section, the test results of three escalators are analyzed. The sixth section is the conclusion.

2. Theoretical Analysis

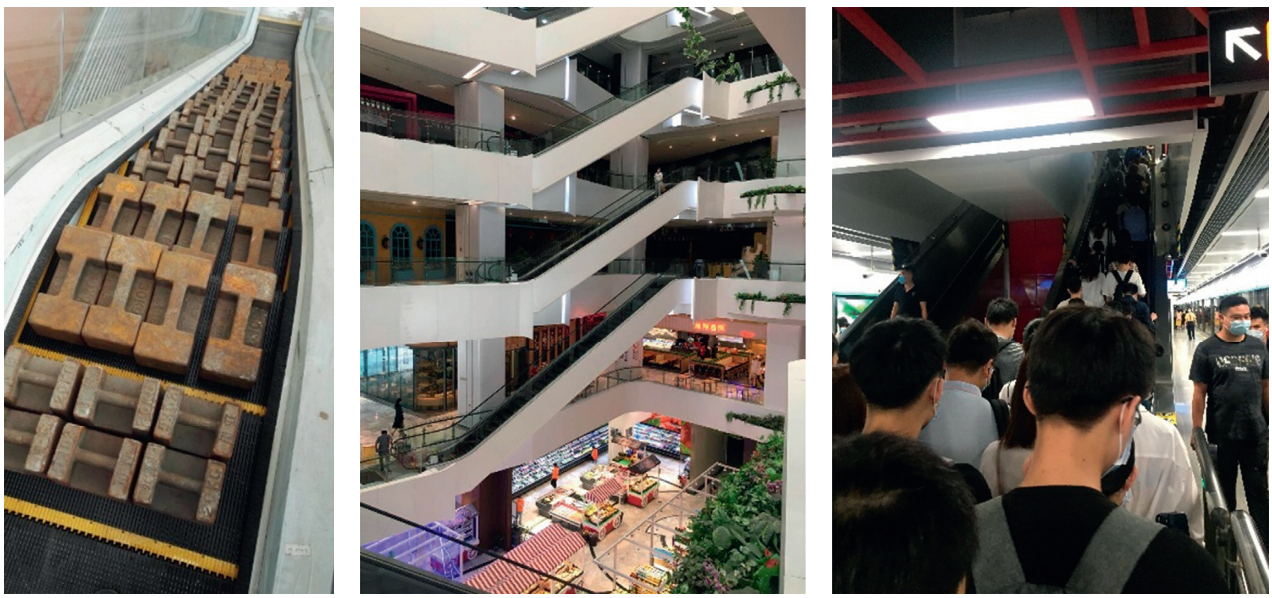
2.1. General Equations of Mechanical Motion. For a mechanical system composed of moving components, the acting force on the component i is F_i, the torque is M_i, the velocity of the point of force is v_i, the angular velocity of the component is ω_i, the velocity of the center of mass is v_{si}, and the moment of inertia of the center of mass is J_{si}, then there is

$$d\left[\sum_{i=1}^{n}\frac{m_i v_{si}^2}{2}+\frac{j_{si}\omega_i^2}{2}\right]=d\left[\sum_{i=1}^{n}\left(F_i v_i\cos\alpha_i \pm M_i\omega_i\right)\right]\cdot \mathrm{d}t, \tag{1}$$

where α_i is the angle between force and velocity, plus or minus sign depends on the direction of the torque M_i acting on the component and the angular velocity ω_i of the component. When they are the same, "+" is taken, and when they are opposite, "−" is taken.

Equation (1) shows that, for complex systems, there are many components and it is difficult to solve. However, single-degree-of-freedom mechanical systems can be simplified by equivalent.

2.2. Equivalent Dynamic Model of Escalator System. According to the mechanical principle, the single-degree-of-freedom mechanism can be reduced to an equivalent component with equivalent mass or equivalent rotational inertia [28]. At this time, the motion law of the equivalent component is the same as that in the mechanism. The condition of the equivalent rotational inertia is that the kinetic energy of the equivalent component with the equivalent rotational inertia (mass) is equal to the kinetic energy of the original mechanical system. Due to the kinetic energy theorem, during mechanical operation, the elemental work done dw by all external forces in any time interval dt should be equal to the increase in dE in kinetic energy of the mechanical system:



(a) (b) (c)

Figure 5: Escalator status: (a) escalator loading test; (b) escalator of mall; (c) status of escalator operation.

$$\mathrm{dw} = \mathrm{d}E. \tag{2}$$

For the escalator system shown in Figure 6, let the angular speed of the step sprocket be ω_1, the mass is m_1, the moment of inertia is J_1, the mass of all the steps is m_2, the speed of the step is v_2, the load mass is m_3, and the speed of load is v_3.

When running with no load, there is

$$\begin{aligned} \mathrm{d}E &= d\left(\frac{J_1\omega_1^2}{2} + \frac{m_2 v_2^2}{2}\right), \\ \mathrm{dw} &= M_1\omega_1 \cdot \mathrm{d}t. \end{aligned} \tag{3}$$

From equation (2),

$$d\left\{\frac{\omega_1^2}{2}\left[J_1 + m_2\left(\frac{v_2}{\omega_1}\right)^2\right]\right\} = M_1\omega_1 \cdot \mathrm{d}t. \tag{4}$$

Make $J_{e0} = J_1 + m_2 \cdot (v_2/\omega_1)^2$ and $M_{e0} = M_1$, where J_{e0} is the moment of inertia equivalent to the step sprocket, M_{e0} is the equivalent torque of the step sprocket, and M_1 is the electromagnetic torque converted to the step sprocket.

When running down with rated load, there is

$$\begin{aligned} \mathrm{d}E &= d\left(\frac{J_1\omega_1^2}{2} + \frac{m_2 v_2^2}{2} + \frac{m_3 v_3^2}{2}\right), \\ \mathrm{dw} &= (m_3 g \sin\alpha \cdot v_3 - M_1\omega_1) \cdot \mathrm{d}t. \end{aligned} \tag{5}$$

Due to $v_2 = v_3 = v$,

$$d\left\{\frac{\omega_1^2}{2}\left[J_1 + m_2\left(\frac{v}{\omega_1}\right)^2 + m_3\left(\frac{v}{\omega_1}\right)^2\right]\right\} = \omega_1\left[\frac{m_3 g \sin\alpha \cdot v}{\omega_1} M_1\right]\mathrm{d}t. \tag{6}$$

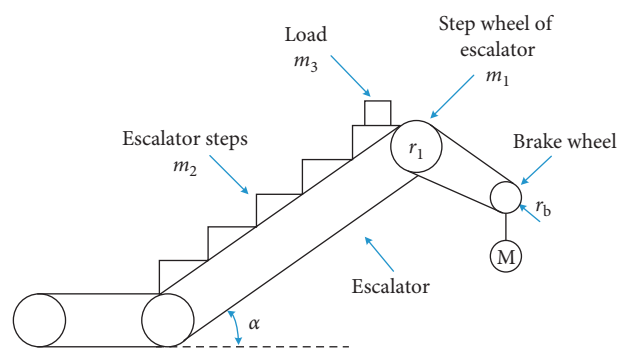


Figure 6: The motion model of escalator.

Make $J_{e1} = J_1 + (m_2 + m_3) \cdot (v/\omega_1)^2$ and $M_{e1} = ((m_3 g \sin\alpha \cdot v/\omega_1) - M_1)$, where J_{e1} is the equivalent moment of inertia of the step sprocket and M_{e1} is the equivalent torque of the step sprocket.

The equivalent moment of inertia converted to the brake wheel is

$$J'_{e1} = J_{e1} \cdot \left(\frac{n_1}{n_b}\right)^2 \cdot \eta. \tag{7}$$

In equation (7), n_1 is the speed of the step sprocket, m/s; n_b is the speed of the brake wheel, m/s; and η is the transmission efficiency.

The torque converted to the brake wheel is

$$M'_{e1} = \frac{M_{e1}}{\eta} \cdot \frac{n_1}{n_b}. \tag{8}$$

On applying brake to the escalator, the braking distance is

$$S = \frac{v^2}{2a}, \tag{9}$$

where v is the escalator running speed, m/s, and a is the deceleration, m/s^2.

The angular deceleration converted to the brake wheel is

$$\varepsilon_b = \frac{a}{R_1} \cdot \frac{n_1}{n_b}, \tag{10}$$

where R_1 is the radius of the step sprocket, m.

Since the braking torque of the brake is constant, set to M_b, there is

$$\varepsilon_b = \frac{M_b - M'_{e1}}{J'_{e1}}. \tag{11}$$

From equations (9)–(11),

$$\begin{aligned} S &= \frac{v^2 \cdot n_1 \cdot J'_{e1}}{2R_1 n_b \left(M_b - M'_{e1}\right)} \\ &= \frac{v^2 \cdot n_1 \cdot \left[J_1 + (m_2 + m_3) \cdot (v/\omega_1)^2\right] \cdot (n_1/n_b)^2}{2R_1 n_b \left[M_b - \left((m_3 g \sin\alpha \cdot v/\omega_1) - M_1\right) \cdot (n_1/n_b)\right]}. \end{aligned} \tag{12}$$

It can be seen from equation (12) that when applying brake, the motor power is disconnected, and the electromagnetic torque is zero. The braking distance is related to the mass of the load. As the load increases, the numerator increases, the denominator decreases, and the braking distance increases rapidly. Although the values such as moment of inertia and braking torque can be measured, they involve many and complicated parameters. For the installation site of the escalator, even some parameters cannot be obtained. Therefore, it is very difficult to calculate the braking distance with rated load and can only be measured by loading. This article attempts to analyze the relationship between the braking distance and the load from another angle to achieve the purpose of predicting the braking distance with rated load. By the analysis of energy changes during braking, a mathematical model of braking distance and load is established based on energy conservation, and the influence of some intermediate details is equivalent to a coefficient. Finally, the braking distance prediction of the escalator with rated load and various loads is realized.

3. Prediction Model of Braking Distance of Escalator

3.1. Energy Analysis of Escalator Braking Process. The brakes of escalator products are currently mechanical; that is, the brake system relies on the friction torque to stop the system that drives the main engine to run. Applying brake to an escalator or moving walkway running at a rated speed is the mechanics to consume all the inertial energies of moving parts of the stairway and the load through the frictional resistance between the brake wheel and the brake shoe and stop it within a distance. Taking an escalator as an example, the energy change during braking is shown in Figure 7.

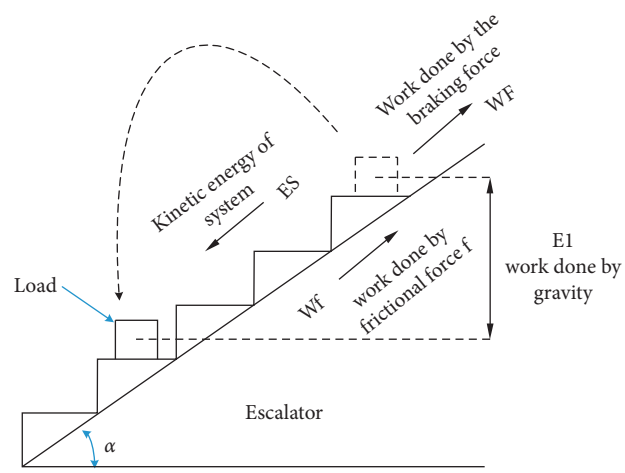


Figure 7: Energy changes during escalator braking.

In Figure 7, the energy of the escalator movement process includes the following: the kinetic energy of the system (including all linear moving parts and rotating moving parts), the work done by the friction of the system, the work done by the gravity of the load, and the work done by the braking force during braking.

3.2. Mathematical Model of Relationship between Braking Distance and Load. The escalator is driven by the motor through the gears to drive the steps and the handrails to make a circular movement. When the fault occurs or the safety switch is activated, the safety circuit is disconnected and the motor loses its power source. At this time, the brake coil is deenergized, and the brake decelerates the brake wheel to stop. During braking, the kinetic energy and potential energy are reduced, the braking force does work, and finally reaches a state of equilibrium.

According to the law of conservation of energy, the sum of the work done by the braking force and the friction force is equal to the reduction of the kinetic energy of the system and the reduction of the potential energy of load. Then,

$$W_{\mathrm{F}} + W_{\mathrm{f}} = E_{\mathrm{S}} + E_{\mathrm{l}}, \tag{13}$$

where W_{F} is the work of the braking force, W_{f} is the work of the friction, E_{S} is the kinetic energy of the system, and E_{l} is the work of gravity of the load.

In order to simplify the calculation, an equivalent method is adopted. Simplify the influence of friction and efficiency into an influence coefficient λ, so as to achieve the equivalent purpose. The core idea is to use the braking force to do work to consume the energy of the system. Two cases of no load and rated load are taken to be considered. Assume that the no load and the rated load have the same influence coefficient. When the escalator is running without load, the work done by the braking force is equal to the reduction of kinetic energy multiplied by the influence factor:

$$FS_0 = \lambda E_0. \tag{14}$$

When the escalator is loaded with a mass of m_1, there are

$$\mathrm{FS}_1 = \lambda\left(\frac{1}{2}m_1 v^2 + E_0 + m_1 g S_1 \sin\alpha\right), \tag{15}$$

where F is the braking force equivalent to the step chain; kinetic energy with no load is $E_0' = (1/2)mv^2 + (1/2)J\omega^2$; m_1 is the load mass; S_1 is the braking distance corresponding to m_1; S_0 is the braking distance with no load; m is the mass of linear moving parts with no load; J is the rotational inertia of the rotating parts with no load; E_0 is the ENKE related to E_0'; that is, the energy consumed in addition to the load energy; and λ is the influence coefficient, which is affected by unknown parameters such as friction and transmission efficiency.

From equations (1) and (2), the relationship between the braking distance and the load is as follows:

$$S_1 = \frac{(1/2)m_1 v^2 S_0 + E_0 S_0}{E_0 - m_1 g \sin\alpha \cdot S_0}. \tag{16}$$

Equations (16) and (12) show the relationship between the braking distance and the load from two different angles. From the comparison between equations (16) and (12), it can be seen that they are similar in the following aspects: when the load increases, the numerator increases and the denominator decreases; the braking distance increases with the increase in load, and the closer to the rated load, the greater the increase in braking distance. However, the parameters to be measured in equation (16) are significantly less than those in equation (12), which shows that using equation (16) to predict the braking distance is more simple and convenient.

3.3. Test and Verification. The braking distance of the escalator under different loads can be obtained by equation (16). Although the ENKE in the model is a fixed value, it is related to many factors, and it is difficult to calculate accurately. In order to simplify the calculation, this paper proposed the back derivation method to obtain the ENKE. And then, substitute it into equation (16) to obtain the braking distance of the corresponding load. This not only avoids the high cost but also avoids the safety hazards of rated load testing.

Principle of back derivation method: for equation (16), braking distance with no load can be measured experimentally. v, g, and α are all fixed values, m_1 is the mass of the load, and E_0 is the ENKE, which is a constant related to the braking force. According to the analysis, when the load mass m_1 is given, the braking distance S_1 can be obtained. Conversely, by testing the value of S_1 under load m_1, the magnitude of the ENKE E_0 can be calculated. If two braking distance tests are conducted under no-load and few-load conditions, respectively, the ENKE under the current conditions can be obtained by equation (16).

After calculating the ENKE by the back derivation method, only the braking distance S_1 and the load m_1 to be loaded are unknown in equation (16). This means that, as long as the load value is given, the braking distance under any load can be obtained. Take 3 different escalators as examples to verify equation (16).

The infrared-based ranging sensor tester is used for the braking distance test of escalator, which is designed independently. The parameters of the instrument are shown in Table 2.

3.3.1. Test Method. Place the instrument on the step, reserve enough braking distance, and start the escalator operation; when the escalator reaches a uniform speed, press the emergency stop button and test the relative distance of the instrument in real time during the braking process. Once the emergency stop signal is detected, the relative distance is recorded once. After the escalator stops, the relative distance is recorded again. The distance between the two measured values is the escalator braking distance. The test site is shown in Figure 8.

Figure 8(a) is the braking distance test when there is no load, and Figure 8(b) is the braking distance test when it is rated load.

The parameters of the three tested escalators are shown in Table 3.

3.4. Verification and Analysis. The test of the three escalators was carried out using the test method of braking distance in Section 3.3. The braking distance of the escalators was measured under the conditions of no load and 25%, 50%, 75%, and 100% rated load, respectively. The measured results of the braking distance when there were different loads are shown in Table 4.

As can be seen from Table 4, for the escalator 1, when the load is 0 kg, 450 kg, and 1800 kg, the measured braking distance is 0.3 m, 0.38 m, and 0.73 m, respectively. As the load increases, the braking distance increases, and when the load exceeds 50% of the rated load, the increase in braking distance increases significantly, which basically increases exponentially. For escalators 2 and 3, the same results as escalator 1 are obtained.

Equation (16) considers the ENKE to be a constant value. To verify, the back derivation method is used. Take escalator 1 as an example. Substitute the corresponding braking distance at 0 kg and at 450 kg into equation (16). The corresponding ENKE is calculated as 3353.0625 J. In the same way, the braking distances corresponding to 0 kg and other loads are substituted into equation (16), and the ENKE corresponding to the various loads calculated is 4194 J, 4291.553571 J, and 4649.023256 J. Similarly, the calculation results of escalators 2 and 3 are shown in Table 4.

It can be seen from the results in Table 4 that the calculated ENKE is inconsistent at different loads. As the load increases, the calculation result of the ENKE also increases. From 25% to 50% rated load stage, the ENKE increases faster; from 50% to 75% rated load stage, the ENKE increases slowly; and from 75% to 100% rated load stage, the ENKE increase rate increases again.

Through theoretical analysis, the ENKE proposed in this paper is related to the braking force. For an escalator, under the condition of a certain braking torque, the ENKE will not change. Then, the corresponding ENKE should be a certain value. However, from the calculation results of the no-load

Table 2: Parameters of the instrument.

Sampling frequency	25 Hz	Resolution	0.1 mm
Measuring range	0.1 ~ 100 m	Transmission interface	Bluetooth
Measurement accuracy	±3 mm	Operating temperature	−10 ~ +50°C

(a) (b)

Figure 8: Braking distance test: (a) with no load; (b) with load.

Table 3: Parameters of tested escalators.

Parameters	Escalator 1	Escalator 2	Escalator 3
Nominal speed	0.5 m/s	0.5 m/s	0.5 m/s
Tilt angle	35°	35°	35°
Lifting height	4.2 m	5.3 m	4.8 m
Nominal width	1000 mm	1000 mm	1000 mm
Rated load	1800 kg	2760 kg	2400 kg
Power	7.5 kw	11 kw	7.5 kw

equivalent kinetic energy in Table 4, it can be seen that the ENKE becomes larger with the increase in the load. That is to say, the energy needs to be consumed by the brake in addition to the energy of the load itself, and the increased energy of the ENKE is noticed. It is caused by an increase in load. This is inconsistent with theoretical analysis. The analysis shows that, as the load increases, the friction and efficiency of the escalator system will change, so the impact coefficient of this model will also change with the increase in the load. The next section will improve the model.

4. Improved Model

4.1. Modification of the Model. According to the test results and theoretical analysis in the previous section, the calculated ENKE is not a constant but increases with the increase in the load. The analysis shows that, due to the existence of factors such as meshing and friction between the gears, the rated conversion efficiency of the work done by the braking force is deviated. The efficiency of the escalator in no-load and loading operation is different. This shows that the coefficients in equations (14) and (15) are not a constant value but change as the load increases.

So, equation (14) is amended as follows when no load:

$$\mathrm{FS}_0 = k_0 E_0. \tag{17}$$

Equation (15) is amended as follows when loading:

$$\mathrm{FS}_1 = k_1\left(\frac{1}{2}m_1 v^2 + E_0 + m_1 g S_1 \sin\alpha\right). \tag{18}$$

Kinetic energy with no load is as follows:

$$E_0' = \frac{1}{2}mv^2 + \frac{1}{2}J\omega^2. \tag{19}$$

Among them, k_0 is the influence coefficient with no load, and k_1 is the influence coefficient with loading, which varies with the load. E_0 is the ENKE, which is related to E_0', that is, the energy consumed in addition to the load energy.

From equations (17) and (18), the expression of the baking distance with load m_1 is

Table 4: Test and calculation results of three escalators.

Tested escalator		Load percentage (%)				
		0	25	50	75	100
Escalator 1 (rated load 1800 kg)	Test result of braking distance (m)	0.3	0.38	0.45	0.58	0.73
	Calculation result of ENKE (J)	—	3353.06	4194.00	4291.55	4649.02
Escalator 2 (rated load 2760 kg)	Test result of braking distance (m)	0.34	0.44	0.54	0.72	0.92
	Calculation result of ENKE (J)	—	6059.34	7369.81	7680.53	8305.68
Escalator 3 (rated load 2400 kg)	Test result of braking distance (m)	0.19	0.23	0.26	0.3	0.34
	Calculation result of ENKE (J)	—	4062.84	5195.78	5662.83	6224.59

$$S_1 = \frac{(1/2)m_1 v^2 S_0 + E_0 S_0}{\lambda_1 E_0 - m_1 g \sin\alpha \cdot S_0}, \tag{20}$$

where $\lambda_1 = (k_0/k_1)$.

Equation (20) is deformed:

$$\lambda_1 = \frac{(1/2)m_1 v^2 S_0 + E_0 S_0 + m_1 g \sin\alpha \cdot S_0 S_1}{S_1 E_0}, \tag{21}$$

$$E_0 = \frac{0.5 m_1 v^2 S_0 + m_1 g \sin\alpha \cdot S_0 \cdot S_1}{\lambda_1 S_1 - S_0}. \tag{22}$$

4.2. Determination of Influence Coefficient and Analysis. For the calculation results of the ENKE of the test escalator in Section 3.4, it is actually assumed $\lambda_1 = (k_0/k_1) = 1$. According to this assumption, substitute the test results of the braking distance of Table 4 into equation (22); the ENKE calculation results of the three escalators can be obtained, as shown in Table 4.

Since the improved mathematical model thinks that the value of λ_1 is not 1, the next step is to calculate the value of λ_1. Take escalator 1 as an example. Assuming that the 25% rated load is used as the reference, $\lambda_1 = 1$, so the ENKE at this time is $E_0 = 3353.0625$ J. Take $E_0 = 3353.0625$ J as the ENKE of the system; that is to say, the ENKE of the system is a constant value 3353 J under any load (values after the decimal point are omitted). Then, the λ_1 value under different loads can be obtained from equation (21), as shown in Table 5.

Similarly, for the escalator 2, substitute the test results of the braking distance in Table 4 into equation 22, and it can be calculated as $E_0 = 6059.34264$. The corresponding calculation results of λ_1 under different loads are shown in Table 6.

As can be seen in Table 6, the 25% rated load is used as the reference $\lambda_1 = 1$. When the load is 50%, 75%, and 100% rated load, the calculated λ_1 is 1.080, 1.141, and 1.239, respectively.

For the escalator 3, substitute the test results of the braking distance in Table 4 into equation (22), and it can be calculated as $E_0 = 4062.8403$ J. The corresponding calculation results of λ_1 under different loads are shown in Table 7.

It can be seen from Table 7 that when the load is 50%, 75%, and 100% rated load, the calculated λ_1 is 1.075, 1.144, and 1.234, respectively.

This is approximately the same as the values of escalators 1 and 2 under the same load. Then, the comparison of the calculation results of the three escalators is shown in Figure 9.

It can be intuitively seen from Figure 9 that the λ_1 value change trend of the three escalators under different loads is very consistent and basically coincides at a specific point. Therefore, it can be obtained that when the load is 25%, $\lambda_1 \approx 1$; when the load is 50%, $\lambda_1 \approx 1.08$; when the load is 75%, $\lambda_1 \approx 1.14$; when the load is 100%, $\lambda_1 \approx 1.234$, as shown in Table 8.

The λ_1 value corresponding to the load not listed in Table 8 is calculated by the piecewise interpolation method.

The test results of the above three escalators show that, with the increase in load, λ_1 shows an increasing trend. But the growth rate decreases during the 50%–75% rated load stage, and the growth rate is faster at the 75%–100% rated load stage. The inflection point appears around 75% of the rated load. According to equation (20), as the load increases, the numerator becomes larger and the denominator becomes smaller. Due to the existence of λ_1, after 75% rated load, the speed at which the denominator becomes smaller is slowed down. This suppresses the excessively rapid increase in the braking distance, and a reasonable braking distance value is obtained.

So far, the improved braking distance prediction model of escalator has been completed, and the influence coefficient has been determined. To verify the improved mathematical model, the next section will test the model and compare it with the measured value.

5. Test Results of Improved Model

Since the determination of the parameters in the improved model is based on the three escalator tests mentioned in Sections 3.3, the tilt angles are all 35°. In order to avoid interference, other three escalators with different models, different angles, and different rated loads were used as examples for testing. The parameters of the escalators are shown in Table 9.

The loading test process is shown in Figure 10.

The actual measurement results of the braking distance are shown in Table 10.

It can be seen from Table 10, for escalator 1, the braking distance with no load is small, and the increment between no-load and 25% rated load is small, so the braking distance increases slowly as the load increases, and the braking distance at rated load is also smaller, only 0.68 m; for escalator 2, the no-load braking distance is larger, and the

Table 5: Calculated result of λ_1 for escalator 1.

Load percentage (%)	25	50	75	100
Equivalent no-load kinetic energy (J)	3353	3353	3353	3353
λ_1	1.000003924	1.08360672	1.135131174	1.227679976

Table 6: Calculated result of λ_1 for escalator 2.

Load percentage (%)	25	50	75	100
Equivalent no-load kinetic energy (J)	6059	6059	6059	6059
λ_1	1.000012852	1.080126793	1.141245873	1.239476107

Table 7: Calculated result of λ_1 for escalator 3.

Load percentage (%)	25	50	75	100
Equivalent no-load kinetic energy (J)	4062	4062	4062	4062
λ_1	1.000035977	1.075147415	1.144503496	1.234880806

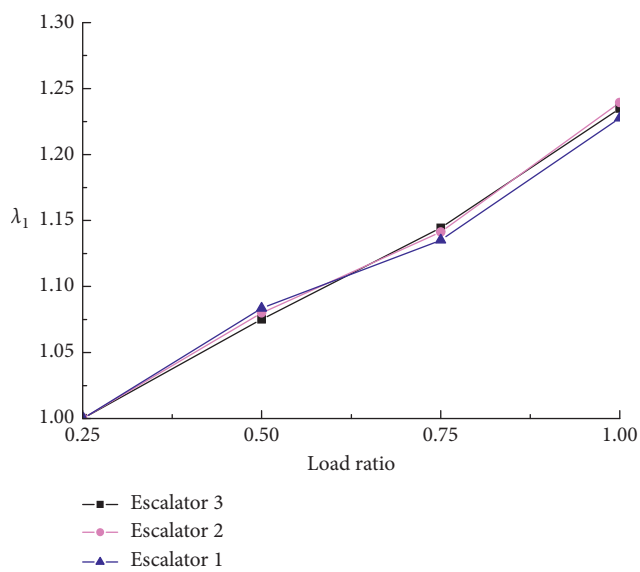


Figure 9: Comparison of λ_1 for three escalators.

Table 8: Values of λ_1.

Load percentage (%)	25	50	75	100
λ_1	1.000	1.080	1.140	1.234

increase in the braking distance between no-load and 25% rated load is larger, so with the increase in load, the braking distance increases rapidly. The braking distance at rated load is 1.09 m, which has exceeded the standard requirements. Therefore, it can be concluded that, as the load increases, the braking distance shows an upward trend. When the braking distance change between no-load and 25% rated load is greater, the braking distance increases faster and is likely to exceed the standard. On the contrary, the growth of the braking distance is relatively slow.

Table 9: Parameters of tested escalators.

Parameters	Escalator 1	Escalator 2	Escalator 3
Nominal speed	0.5 m/s	0.5 m/s	0.5 m/s
Tilt angle	30°	35°	35°
Lifting height	3.6 m	5.3 m	5.0 m
Nominal width	1000 mm	1000 mm	1000 mm
Rated load	2000 kg	2880 kg	2640 kg
Power	7.5 kw	11.0 kw	8.0 kw

According to the back derivation method in Section 3.3, the actual measurement results of the no-load and 25% rated load in Table 10 are substituted into equation (22). And the ENKE of each escalator is calculated under the current braking force, as shown in Table 11.

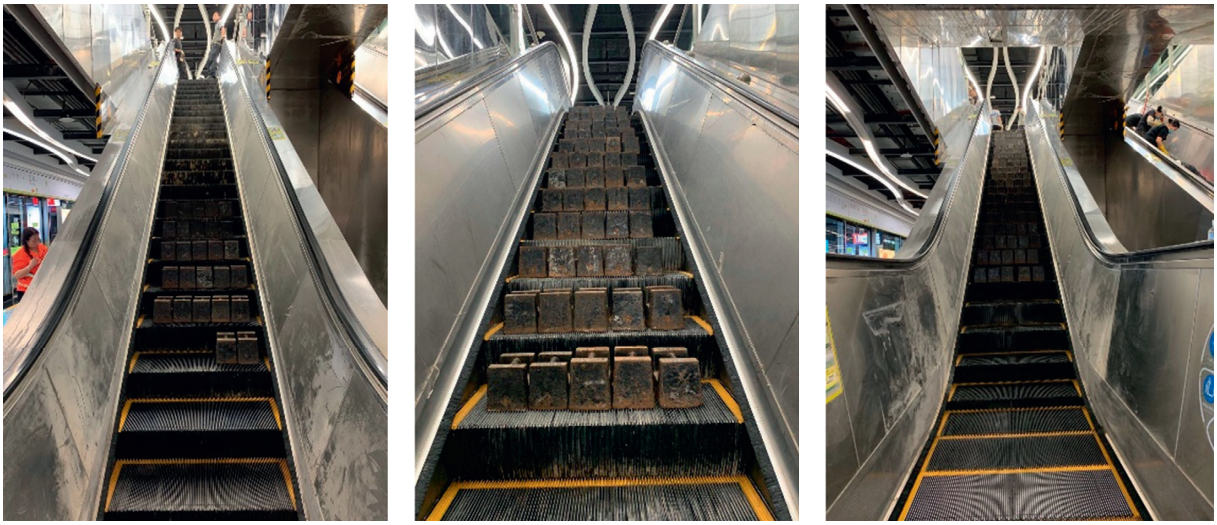
The ENKE of each escalator calculated in Table 11 is substituted into equation (20), and the predicted results of the braking distance under different loads of each escalator are calculated, as shown in Table 12.

It can be seen from Table 12 that, for an escalator of 0.5 m/s, when the braking distance with no load is close to 0.2 m, the braking distance with rated load is predicted to be 0.67 m. And there is space from the standard requirement of 1 m; when the braking distance with no load exceeds 0.4 m, the predicted value of the braking distance with rated load is basically close to the upper limit of the standard or even exceeds the standard. Relevant measures need to be taken to reduce the braking distance to ensure the safe of the escalator.

6. Analysis and Discussion

According to the test results in Section 5 above, the comparison between the predicted value of the improved braking distance prediction model and the measured value is shown in Figure 11.

By combining Tables 10 and 12 and Figure 11, it can be seen that the difference between the predicted value and the



(a) (b) (c)

Figure 10: Load test of braking distance.

Table 10: Actual measurement results of braking distance.

Load percentage (%)	0	25	50	75	100
Braking distance of escalator 1 (m)	0.23	0.3	0.38	0.5	0.68
Braking distance of escalator 2 (m)	0.45	0.57	0.7	0.88	1.09
Braking distance of escalator 3 (m)	0.43	0.54	0.65	0.81	1.0

Table 11: Calculated values of equivalent no-load kinetic energy.

Escalator no.	Escalator 1	Escalator 2	Escalator 3
Equivalent no-load kinetic energy (J)	2620.35	10441.08	9879.85

Table 12: Predicted value of braking distance for 3 escalators.

Load percentage	0	25%	50%	75%	100%
Predicted braking distance of escalator 1 (m)	0.230	0.300	0.370	0.498	0.674
Predicted braking distance of escalator 2 (m)	0.450	0.570	0.681	0.873	1.112
Predicted braking distance of escalator 3 (m)	0.430	0.540	0.637	0.803	0.997

measured value is 0.022 m at the maximum, and the maximum error is 2.7%. It can meet the needs of engineering measurement. At the same time, the error we are more concerned about at 100% of the rated load is smaller, the minimum difference is only 0.003 m, and the minimum error is 0.3%. This means that the predicted value is closer to the actual value at 100% of the rated load.

It can be seen in Figure 11 that, before 50% rated load, the slope of the curve of braking distance with load is smaller; after that, the slope of the curve of braking distance with load is larger, which means that, from 50% rated load, as the load increases, the braking distance increases faster. The analysis shows that, as the load increases, the kinetic energy of the escalator system continues to increase. At 50% rated load, on applying brake, due to the large impulse force, the critical point of friction is reached, causing the friction of the system to become sliding friction. The frictional force is reduced so that the braking distance increases more obviously.

Furthermore, in order to visually show the influence of the braking force on the braking distance, the braking force is adjusted on the escalator 1. The braking force is increased first, and then, the braking force is reduced. The braking distance test method in Section 3.3 has been used for testing. The measured results of the braking distance before and after adjustment are shown in Table 13.

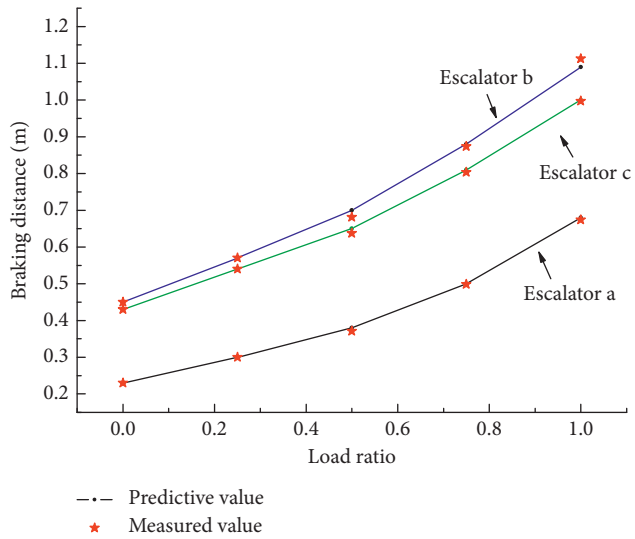


Figure 11: Measured and predicted values of three escalators.

Table 13: Test results before and after adjustment.

Load percentage (%)	0	25	50	75	100	Remarks
Braking distance of escalator (m)	0.23	0.3	0.38	0.5	0.68	Before adjustment
Braking distance of escalator (m)	0.18	0.24	0.32	0.43	0.62	Increase the braking force
Braking distance of escalator (m)	0.26	0.34	0.42	0.6	0.8	Decrease the braking force

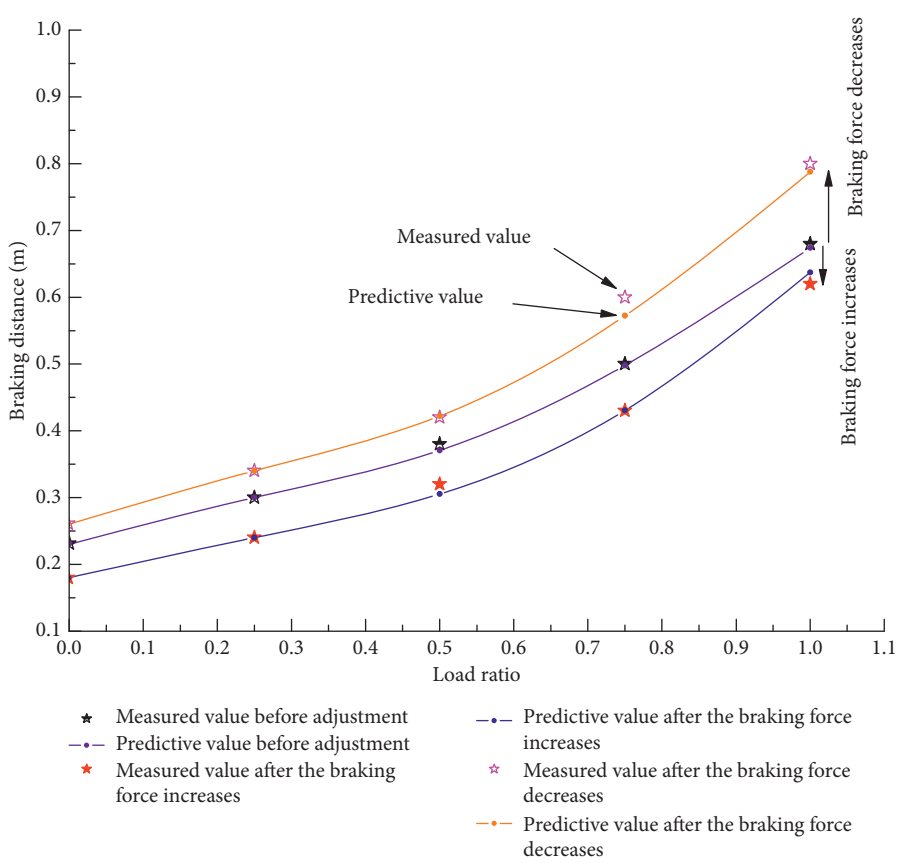


Figure 12: Comparison of braking distance before and after braking force adjustment.

According to the back derivation method in Section 3.3, the actual measurement results of the no-load and 25% rated load in Table 13 are substituted into equation (22), respectively. The value of ENKE is calculated. And then, the predicted value of the braking distance can be obtained, respectively, according to equation (20) before and after adjustment. The comparison between the predicted value and the measured value is shown in Figure 12.

It can be seen from Figure 12 that the test value and the predicted value of the braking distance before and after the braking force adjustment are basically the same, indicating that the improved prediction model can meet the needs of engineering measurement. Before the braking force adjustment, the braking distance with no load is 0.23 m. After the braking force increases, the braking distance with no load is 0.18 m, indicating that the braking force will decrease when the braking force increases; after the braking force decreases, the braking distance with no load is 0.26 m, indicating that the braking force will increase if the braking force decreases. The same phenomenon happened when rated loaded. Further analysis, from the braking distance test results of Table 13 and equation (22), the ENKE can be calculated to be 2620 J, before the adjustment of the braking force. After the brake spring is tightened, that is, the braking force increases. Then, the result is that the braking distance becomes smaller. Correspondingly, from equation (22), the ENKE is 2223 J, which also becomes smaller. After the brake spring is loosened, it means that the braking force is reduced. Then, the result is that the braking distance becomes larger. Correspondingly, from equation (22), the ENKE is 2910 J, which also becomes larger. It can be concluded that the ENKE of an escalator with a large braking force is small; that is, the ENKE that needs to be consumed is small. In addition, the change in braking distance between no-load and 25% rated load is large, indicating that the braking force is small, and the braking distance increases faster as the load increases.

7. Conclusion

In this paper, the relationship model between the braking distance and the load of the escalator is derived through the analysis of the force and energy changes in the escalator during braking. After experimental verification, it was found that the influence coefficients were inconsistent under no-load and loaded conditions, and the model was revised accordingly, and an improved model was obtained. It realizes braking distance predicting various loads for escalator under light load test conditions, greatly reducing the safety hazards and test costs on traditional testing of braking distance under rated load and improving test efficiency:

(1) The prediction model before improvement considers that the influence coefficients are equal whether no load or loaded. However, it has been verified by experiments that the influence coefficient increases nonlinearly with the increase in the load. The improved prediction model revises the influence coefficient and proposes the change law of the influence coefficient on the braking distance. The results show that the maximum error between the braking distance predicted by the improved model and the braking distance measured by the loading method is 2.3%, which meets the test error of engineering application.

(2) As the load increases, the braking distance is on the rise. When the difference of braking distance between no-load and 25% rated load is large, the braking distance increases faster and is likely to exceed the standard. On the contrary, the growth of the braking distance is relatively slow.

(3) The braking distance test is based on the power-off time, including the electrical and mechanical delay time of the brake. And the load influence coefficient is also obtained on this basis. If the braking distance test starts from braking, the load influence coefficient needs to be recalculated.

(4) The introduction of ENKE simplifies the model and ignores the influence of some intermediate quantities, which simplifies the calculation. At the same time, a back derivation method is proposed to calculate the ENKE, which avoids the complexity and uncertainty of the ENKE calculation.

(5) The improved prediction model shows that the escalator with a large braking force has a small ENKE; that is to say, for the braking force under this condition, the escalator is easy to stop. Conversely, the escalator with a small braking force has a large ENKE, so the escalator is not easy to stop, which is consistent with the actual situation.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded by the Guangzhou Market Supervision Administration, China (no. 2020kj26), and Guangzhou Science and Technology Bureau, China (no. 202002030121).

References

[1] Guangdong Administration for Market Regulation, *Notice of the Guangdong Administration for Maket Regulation on the Safety Status of Special Equipment in Guangdong Province in 2019 EB/OL*, 2020. http://amr.gd.gov.cn/.

[2] S. Uimonen, T. Tukia, J. Ekstrom et al., "A machine learning approach to modelling escalator demand response," *Engineering Applications of Artificial Intelligence*, vol. 90, Article ID 103521, 2020.

[3] Z. Zhou, Y. Zi, J. Chen et al., "Hazard analysis for escalator emergency braking system via system safety analysis method based on STAM," *Applied Sciences*, vol. 9, no. 21, Article ID 4530, 2019.

[4] K. Xie and Z. Liu, "Factors influencing escalator-related incidents in China: a systematic analysis using ism-dematel method," *International Journal of Environmental Research and Public Health*, vol. 16, no. 14, 2019.

[5] N. Rohatgi, K. Mehta, P. Sarkar, and T. C. Michael, "Emergency braking mechanism for an elevator using hydraulic and pneumatic actuation," *International Journal of Reliability and Safety*, vol. 13, no. 1-2, pp. 125–137, 2019.

[6] K. M. Mishra and K. Huhtala, "Elevator fault detection using profile extraction and deep autoencoder feature extraction for acceleration and magnetic signals," *Applied Sciences*, vol. 9, no. 15, p. 2990, 2019.

[7] Y. Xing, S. Dissanayake, J. Lu, S. Long, and Y. Lou, "An analysis of escalator-related injuries in metro stations in China, 2013–2015," *Accident Analysis & Prevention*, vol. 122, pp. 332–341, 2019.

[8] Z. Chen and J. Xian, "Escalator-related injuries against preschoolers: an in-depth investigation in Guangdong province, China," *Injury Prevention*, vol. 22, no. 2, 2016.

[9] Z. Chen and J. Huang, "Escalator-related injuries against preschoolers: an in-depth investigation in Shanxi province, China," *Injury Prevention*, vol. 24, no. 2, 2018.

[10] A. Algin, U. Gulacti, M. O. Erdogan, I. Tayfur, K. Yusufoglu, and U. Lok, "Escalator-related injuries in one of the deepest subway stations in Europe," *Annals of Saudi Medicine*, vol. 39, no. 2, pp. 112–117, 2019.

[11] S. S. M. Gob, S. L. Chong, and A. Tyebally, "Danger in shopping centres—a study on escalator-related injuries in children in Singapore," *Annals Academy of Medicine Singapore*, vol. 47, no. 8, pp. 353–356, 2018.

[12] Q. Peng, Z. Li, H. Yuan, G. Huang, S. Li, and X. Sun, "A model-based unloaded test method for analysis of braking capacity of elevator brake," *Advances in Materials Science and Engineering*, vol. 2018, Article ID 8047490, 10 pages, 2018.

[13] Y. Xing, S. Chen, S. Zhu et al., "Analysis factors that influence escalator-related injuries in metro stations based on bayesian networks: a case study in China," *International Journal of Environmental Research and Public Health*, vol. 17, no. 2, p. 481, 2020.

[14] Q. Peng, A. Jiang, H. Yuan, G. Huang, S. He, and S. Li, "Study on theoretical model and test method of vertical vibration of elevator traction system," *Mathematical Problems in Engineering*, vol. 2020, Article ID 8518024, 12 pages, 2020.

[15] J. Sun, M. Yuan, and Z. Lin, "Classification and risk prevention of escalator accidents—A study on risk prevention and evaluation of escalator accidents in escalator," *Journal of Beijing Vocational College of Politics and Law*, vol. 2, pp. 100–105, 2017.

[16] S. M. Kalikate, S. R. Patil, and S. M. Sawant, "Simulation-based estimation of an automotive magnetorheological brake system performance," *Journal of Advanced Research*, vol. 14, pp. 43–51, 2018.

[17] E. Durak and H. A. Yurtseven, "Experimental study of the tribological properties of an elevator's brake linings," *Industrial Lubrication and Tribology*, vol. 68, no. 6, pp. 683–688, 2016.

[18] H. Lü, Z. Cai, Q. Feng, W.-B. Shangguan, and D. Yu, "An improved method for fuzzy-interval uncertainty analysis and its application in brake instability study," *Computer Methods in Applied Mechanics and Engineering*, vol. 342, pp. 142–160, 2018.

[19] Z. Wang, L. Qin, and S. Xiuya, "Analysis of power supply immunity based on heavy-duty escalator brakes," *China Elevator*, vol. 15, no. 30, pp. 9–12, 2019.

[20] J. Bao, Z. Li, and D. Hu, "Frictional performance and temperature rise of a mining nonasbestos brake material during emergency braking," *Advances in Materials Science and Engineering*, vol. 2015, Article ID 867549, 7 pages, 2015.

[21] C. J. Park and G. Gschwendtner, "Braking performance analysis of an escalator system using multibody dynamics simulation technology," *Journal of Mechanical Science and Technology*, vol. 29, no. 7, pp. 2645–2651, 2015.

[22] G. Pan and L. Chen, "Impact analysis of brake pad backplate structure and friction lining material on disc-brake noise," *Advances in Materials Science and Engineering*, vol. 2018, Article ID 7093978, 9 pages, 2018.

[23] General Administration of Quality Supervision, *Inspection and Quarantine Code for Elevator Supervision, Inspection and Periodic Inspection-Escalators and Moving walkwayS: TSG T 7005-2012*, General Administration of Quality Supervision, Beijing, China, 2012.

[24] W. Wang and J. Lu, "Calculation of braking capacity of escalators and travelator," *Mechanical Design*, no. 5, pp. 18-19, 2006.

[25] X. Pan, "The stopping distance of escalators," *China Special Equipment Safety*, vol. 32, no. 10, pp. 27–29, 2012.

[26] A. Hu, "A brief talk on a new method to establish a mathematical model of escalator braking distance," *Electromechanical Information*, no. 24, pp. 51-52, 2014.

[27] Z. Liu, "A brief talk on calculation of braking distance of escalator and travelator. Modern manufacturing technology and equipment," *Lack of Roll*, no. 5, pp. 63-64, 2019.

[28] L. Pu, G. Chen, and L. Wu, *Mechanical Design*, Higher Education Press, Beijing, China, 2019.