Parallel Random Search Algorithm for Large-Scale Constrained Pseudo-Boolean Optimization Problems

L.A. Kazakovtsev
Department of Information Technologies, Siberian State Aerospace University, Krasnoyarsk, Russia

Abstract - Random search methods are successfully implemented for variety of discrete optimization NP-hard problems when any exact solution approaches cannot be implemented due to large computational demands. Initially designed for unconstrained optimization, the probability changing method gives an approximate solution for various linear and non-linear pseudo-Boolean optimization problems with constraints. Although, in case of large-scale problems, the computational demands are also significant and the precision of the result depends on the spent time. For constrained optimization, the search of any feasible solution may take significant computational resources.

In this paper, we consider an approach to developing parallel versions of the algorithms based on the modified probability changing method for constrained pseudo-Boolean optimization. Optimization algorithms are adapted for the systems with shared memory (OpenMP) and cluster systems (MPI). The parallel efficiency is estimated for the large-scale non-linear pseudo-Boolean optimization problems with linear constraints and travelling salesman problem.

Keywords: Discrete optimization, parallel computing, MPI

1 Large-scale pseudo-Boolean optimization problem

Most exact solution approaches to the problem of discrete (combinatorial) optimization (knapsack problem, the traveling salesman problem etc.) are based on branch-and-bound method (tree search) [1,2,5]. Unfortunately, most of such problems are in the complexity class NP-hard and require searching a tree of the exponential size and even parallelized versions of such algorithms do not allow us to solve some large-scale pseudo-Boolean optimization problems in acceptable time without significant simplification of the initial problem.

The heuristic random search methods do not guarantee any exact solution but random search methods are statistically optimal. I.e. the percent of the problems solved “almost optimal” grows with the increase of the problem dimension [1].

Let’s consider the problem:

\[ F|X| = \left(\sum_{i=1}^{N-V} a_i x_i \right) \left(\sum_{i=1}^{N-V} c_i x_i \right) \rightarrow \text{max}; \]

\[ \sum_{i=1}^{N-V} b_{ik} x_i \leq B_k \quad \forall 1 \leq k \leq N_{\text{constr}} \]

(1) \quad (2)

Here, \( a_i, c_i, b_{ik} | 1 \leq i \leq N \cdot V \) are some constants, \( x_i (1 \leq i \leq N) \) are Boolean variables. The objective function is non-linear.

The real large-scale problems have sometimes millions of variables. For example, the problem of assortment planning of the retail trade company [6] may include thousands goods names to be selected which can be shipped from hundreds suppliers and have 3-10 variants of retail price. In general, problems of such kind can be solved only with random search algorithms.

Being initially designed to solve the unconstrained optimization problems, the probability changing method (MIVER) is a random search method organized by the following common scheme [1,2].

1. \( k=0 \), the starting values of the probabilities \( P(p_i=p_{i0}, p_{i1}, ..., p_{ik}) \) are assigned where \( p_{i0}=P(x=1) \). Correct setting of the the starting probabilities is a very significant question for the constrained optimization problems.

2. With probability distributions defined by the vector \( P_i \), we generate a set of the independent random numbers \( x_{i0} \).

3. The function values in these points are calculated: \( F(x_{i0}) \).

4. Some function values from the set \( F(x_{i0}) \) and corresponding points \( x_{i0} \) are picked out (for example, point with maximum and minimum values).

5. On the basis of results in item 4, vector \( P_i \) is modified.

6. \( k=k+1 \), if \( k<R \) then go to 2. This stop condition may differ.

7. Otherwise, stop.

To be implemented for the problems like (1-3), this method has to be modified. The modified version of the variant probability method, offered in [6,7] allows us to solve large-scale problems with dimensions up to millions of Boolean variables.

In case of the large-scale problems, even the calculation of the linear objective function takes significant computational resources. The number of constrains usually also grows with the increase of the problem dimension. So, the calculation of the objective function and the constrains is a very large computational problem if it is repeated lot of times.
That is why, the distribution of the computational tasks between the parallel processors or cluster nodes is very important.

In this paper, we do not consider greedy search algorithms (which are deterministic or also randomized [8, 10]) though they are often used to improve the results of the random search methods as the final step of them.

Also, we do not consider the parallel genetic algorithms [11] though some approaches offered for genetic algorithms may be implemented for random search algorithms parallelization.

Here, we offer an approach of adaptation of the existing programs realizing the random search methods of constrained pseudo-Boolean optimization to be implemented in the parallel systems.

2 Serial version of the algorithm and its parallelization with OpenMP

The scheme of the algorithm for the serial systems is shown in Figure 1, variant 1.
At the step of initialization, all the variables \( p \) (components of the probability vector \( P \)) are set to their initial values \((0 \leq p < 1 \ \forall \ p \in P)\). Since our algorithm realizes the method of constrained optimization, the initial value of the probability variables is sometimes very important. Then, we generate the vectors \( X \) of optimized boolean variables. In our case of constrained problem of knapsack type, the large values of vector \( P \) components generate the values of \( X \) which are out of the feasible solutions area due to the constraints (9,10). Due to the constraints (3), the optimal initial values of the vector \( P \) components do not exceed \( 1/\left(\sum_{i=1}^{N} a_i\right) \) [7].

We set the initial values of the vector \( P \) to \( 1/\left(\sum_{i=1}^{N} a_i\right) \) but we have to reduce this value if several starts of our algorithm give us no results in feasible solutions area. This process is not illustrated in the Figure 1 for the simplicity.

Instead of maximum number of steps (\( M \) in Figure 1), we can use the maximum run time as the stop condition. In some cases, it is reasonable to use the maximum number of steps which do not give us the result better than previous ones as the stop condition.

In the cycle \((i=1,N)\), we generate the set of \( N \) vectors \( X_0 \) in accordance with the probability vector \( P \). Then, the objective function is calculated for each \( X_0 \). Also, we calculate the values of the left parts of our constraints, and introduce the second objective function \( f^P \) (penalty function):

\[
 f_{ki} = F_0 P \left| X \right| = C_{\text{penalty}} \sum_{k=1}^{N} F_{pk} \left| X \right| ; \quad (4)
\]

\[
 f_P^P = F_{pk} \left| X \right| = \sum_{i=1}^{N} \sum_{j=1}^{N} b_{ik} x_{kj}; \quad (5)
\]

Here, \( C_{\text{penalty}} \) is some coefficient. If some estimation of the maximum value of the objective function is available, we can use it as the value of the coefficient \( C_{\text{penalty}} \). For linear objective functions:

\[
 C_{\text{penalty}} = \sum_{i=1}^{N} \left| a_i \right| . \quad (6)
\]

The modified objective function \( f^P \) is the sum of the objective function and penalty.

When all the values of \( f_0 \) and \( f^P \) are calculated, we choose the best (maximum) value for \( f^P \) and, if there were the variants of the vector \( X_0 \) which satisfy all the constraints \( f^P = 0 \) then we choose the maximum value of the objective function \( f_0 \) for that variants of the vector \( X_0 \).

In case of the constrained optimization, we have to use a special method of adaptation of the vector \( P \). The solution of different practical tasks shows the best result if we use the multiplicative adaptation with rollback procedure [2,6]. In this case, the components of the vector \( P \) are never set to the value of 0 or 1 which may cause that all the further generations of the \( X \) vector have the same value of the corresponding component which does not give the feasible solution due to constrains (3).

\[
 p_{k,i} = \begin{cases} 
 p_{k-1,i} \cdot d, & x_{ij}^{\text{max}} = 1 \land x_{ij}^{\text{min}} = 0 \land p_{k-1,i} < 0.5, \\
 1 - \frac{1 - p_{k-1,i}}{d}, & x_{ij}^{\text{max}} = 1 \land x_{ij}^{\text{min}} = 0 \land p_{k-1,i} \geq 0.5, \\
 p_{k-1,i}, & x_{ij}^{\text{max}} = 0 \land x_{ij}^{\text{min}} = 1 \land p_{k-1,i} < 0.5, \\
 1 - (1 - p_{k-1,i}) \cdot d, & x_{ij}^{\text{max}} = 0 \land x_{ij}^{\text{min}} = 1 \land p_{k-1,i} \geq 0.5, \\
 p_{k-1,i}, & x_{ij}^{\text{max}} = x_{ij}^{\text{min}}. 
\end{cases} \quad (7)
\]

Here, \( d \) is the adaptation coefficient. In case of multiplicative adaptation, it does not depend on the step number \( k \). In this case, the absolute value of adaptation step depends on the corresponding value of \( p_{0,i} \).

After several steps, the values of \( P \) vector elements are close to 0 or 1 and the decrease of the adaptation step \( (d) \) results in generation of the similar vector \( X \) exemplars which correspond to some local maximum. The rollback procedure is helpful to avoid that situation. It sets the values of \( P \) vector and its adaptation step \( d \) to initial (or other) values. In simplest case, rollback is performed after several steps which do not improve the best objective function value.

The best results are demonstrated with methods of partial rollback procedure which change some part of \( P \) vector components or change all the components so that their new values depend on previous results. We can use the following rollback formula:

\[
 p_{kj} = (p_{kj} + q_k p_0)/(1 + q_k), \text{ if } p_{kj} < p_0. \quad (8)
\]

Here, \( p_0 \) is the initial value of the probability. The coefficient \( q_k \) may be constant or vary depending on the results of previous steps. For example, it may depend on the quantity of the steps which do not improve the maximum result \( s_m \).

\[
 q_k = w / s_m. \quad (9)
\]

The weight coefficient \( w \) has to be chosen experimentally.

In case of constrained optimization, the choice of the initial value \( p_0 \) can be very important. In some cases, the incorrect initial value causes the generation of \( X \) vector sets that lay out of the feasible solutions area. After several steps, the penalty function minimization process results in adaptation of \( P \) vector which allows to generate \( X \) vector exemplars that satisfy our constraint conditions. But the rollback procedure returns the probability vector to its initial (usually incorrect) value. Let's consider a simplest example. Let our problem have only one constraint like that:

\[
b_1 x_1 + b_2 x_2 + b_3 x_3 + ... + b_D x_D < B. \quad (10)
\]

Here, \( D \) is the dimension of our problem \( (D=N) \).
The elements of the vector $X$ at the first steps are generated so that it is set to the value of 1 with probability $p_0$. The expectation of the left part of (10) is

$$M = p_0b_1 + p_0b_2 + p_0b_3 + \ldots + p_0b_0 = (b_1 + b_2 + b_3 + \ldots + b_0) p_0.$$  

(11)

The maximum values of the objective function are usually achieved at the points where the condition is barely satisfied $(b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n \leq B)$ in our case). Therefore, the optimal value of the initial probability is

$$p_0 = B/(b_1 + b_2 + b_3 + \ldots + b_0).$$  

(12)

The negative effect of the rollback procedure can be reduced by adaptation of the initial value of $p_0$:

$$p_{0k} = (p_{1k-1} + p_{2k-1} + \ldots + p_{0k-1})/D.$$  

(13)

This version of rollback procedure is most actual in case of partial rollback which performed at each step of our algorithm. The implementation of this kind of adaptation & rollback procedure is described below for the parallel version of the optimization algorithm for systems with no-remote memory access (clusters).

The adaptation of our algorithm for multiprocessor systems with shared memory can be performed by the parallel generation of the exemplars of the $X$ vector and their estimation. The scheme of this version of our algorithm is shown in Figure 1, variant 2. If our system has $N_p$ processors, the cycle of generation of $N$ exemplars of the vector $X$ can be divided between the processors. Each processor has to generate $N/N_p$ exemplars of the vector $X$ and calculate the value of the objective function, left parts of the constraint conditions and calculate the modified objective function values.

Organizing of the parallel thread takes significant computational expenses. In [14], they estimate that expenses as 1000 operations of real number division. The experiments at 4-processor system with linear 100-dimension problem (105 constraints) show that the parallel version runs 2.8 times faster than the serial one.

### 3 Version for clusters

Our experiments [6, 7] with adaptation of random search discrete optimization algorithms [1, 2] for parallel execution gives rather effective results with use of multiprocessor systems with shared memory and OpenMP [14]. But the intensive data interchange in running mechanisms of MPI reduce the efficiency of the realization of the same algorithms for PVM and MPI clusters when the number of the nodes increases. That issue makes the usage of the computing cluster ineffective even in comparison with low-cost multi-core systems.

For testing purposes, we used the following form of the objective function and constraints:

$$F\{X\} = \sum_{i=1}^{N} \sum_{j=1}^{V} a_{ij}x_{ij} \sum_{i=1}^{N} \left(1 - c_{ij} \sum_{j=1}^{V} x_{ij}\right) \rightarrow \text{max};$$  

(14)

with constraints:

\[
\begin{align*}
\sum_{i=1}^{N} \sum_{j=1}^{V} b_{ij}x_{ij} & \leq B_1; \\
\sum_{i=1}^{N} \sum_{j=1}^{V} b_{ij}x_{ij} & \leq B_2; \\
\ldots \\
\sum_{i=1}^{N} \sum_{j=1}^{V} b_{ij}x_{ij} & \leq B_{N_{\text{con}}}; \\
\sum_{j=1}^{V} x_{ij} & \leq 1 \forall 1 \leq i \leq N.
\end{align*}
\]

(15)

The coefficients in the constraints in the test set are generated so that the feasible solutions set is 1000-1000000 smaller than $2^N$ (depends on $N$ and randomly selected values). The algorithm replaces the whole set of constraints with penalty function [2, 6, 7]. Problem of a traveling salesman [1, 2, 11]:

$$F\{X\} = -\sum_{i=1}^{N} a_{ij}x_{ij} \rightarrow \text{max};$$  

(17)

The only constraint in this case is that the matrix of Boolean variables $X$ must describe the adjacency matrix of Hamiltonian graph. This cycle must be the only path in that graph. Here, $a_{ij}$ is the distance (cost, time, etc.) between $i$-th and $j$-th places (cities), $a_{ij} = 0 \forall i, a_{ij} = a_{ji} \forall i,j$. If $x_{ij} = 1$ then it means that the salesman must go from $i$-th place to the $j$-th one.

We tried to solve the optimization problems in Argo cluster (SISSA, Trieste, Italy). The program code was adapted for MPI running. Algorithm designed for OpenMP parallelizing (see Figure 1, variant 2) was implemented for parallelizing via MPI. MPI synchronizing procedure was used at the step “Synchronization”. Following this flowchart gave negative results in comparison even with a single node (8 nodes, 2 processes per node, 50 variants of $X$ sets in each generation on each node, 10000 Boolean variables, 11000 constraints).

Analysis of the problem showed that at the step of synchronization, each node spent approximately 12% of time waiting for the other nodes to finish the calculation, data interchange between nodes in this case is rather intensive (10000 Boolean variables had to be sent from each node, then, each one had to receive 10000 real numbers of the probability vector after each step).

Reconstruction of the algorithm so that each node performs $N$ steps (100-500 for that scale of the problem) separately, and comparing the results after the $N$-th step, choosing the best ones by master nodes and broadcasting new probability vector gave much better results. But, in this case, each node had to spend up to 43% of time (!) waiting for the others to complete their calculations.

Another way to organize the parallel execution of the search processes with the maximum independence from each other is to organize multiple starts of the algorithm at all the nodes. The algorithm starts at the nodes with the same or different initial parameters (for example, different initial values of the probability vector elements). These
multiple simultaneous starts are performed instead of the rollback procedure in the serial version of the optimization algorithm. Each node starts the cycle of the random $X$ vector generation with probability vector adaptation. If the algorithm does not improve the best objective function value during several steps, the rollback procedure is performed. When all the nodes reach the stop condition, the results of all the nodes are compared to figure out the best one as the final result of the optimization problem.

In its simplest form, the above approach does not need any modification of the implemented software. The serial version of the algorithm runs at all the nodes simultaneously and the results of the nodes are then compared by the operator or by the special node.

The simultaneous execution of the serial version of the algorithm improves the results of the calculations insignificantly due to the similar behavior of the algorithm at all the nodes. The comparison of the $P$ vector values from different nodes after several steps shows that the difference between them is reduced with each following step and the nodes generate the new exemplars of the $X$ vectors in the very close area. In case of the algorithm with no average probability adaptation, the simultaneous execution of the serial version of the algorithm with different initial probability value gives the results much better than the serial algorithm executed at the single node. But in most cases, the version of the optimization algorithm with the average probability adaptation executed at the single node shows even better results. So, our approach must support the average probability adaptation for the optimization algorithms executed at different nodes (13).

---

**Figure 2. Parallel algorithm description**

a) Initialization (at each node), MPI_INIT
b) Generating N random sets of variables $X$, estimating $F(X)$ and constraints (at each node)
c) Choosing the best and the worst $X$ sets, adaptation of the probability vector $P$ (at each node separately)
d) if I am the master node then:
   - d1) check if any data transfer started at the previous steps has been completed
     - d1.1) if so, then if it was the maximum result reached by the k-th node, check if this result is better than the global result
     - d1.1.1) if so, send message “OK” to the k-th node
     - d1.1.2) if not so, send “NOT OK” to the k-th node and send the actual global maximum and the corresponding $X$ vector to the k-th vector
   - d1.2) if the completed data transfer contains the best vector $X$
     - d1.2.1) then recalculate $F(X)$ and if it is better than the global maximum, assume that this is the new global maximum

d2) if there are any messages from the other nodes waiting to be received and no receiving processes are in progress, start receiving (non-blocking, continue processing)
d3) check if the stop conditions are reached. If so, send “STOP” message to all the processes, wait for all of them to receive it and then MPI_FINALIZE, stop
e) if I am not the master node and I have already started any sending process then
   - e1) check if this process has been completed
   - e2) if so, start receiving of the answer from the master node
f) if I am not the master node and I am waiting for a response from the master then
   - f1) check if the master node has started sending anything
   - f2) if so, start receiving (non-blocking)
g) if I am not the master node and I have already started receiving smth. from the master node then
   - g1) if the receiving has been completed then
     - g1.1) if it was “OK” message, start sending my best $X$ vector (non-blocking)
     - g1.2) if it was “NOT OK” message, start receiving global best $X$ vector (non blocking)
     - g1.3) if it was the global best $X$ vector, recalculate $F(X)$ and, if it is better than my own maximum, partially restart local process to perform further search around the global maximum
     - g1.4) if it was the “STOP” message, then stop, MPI_FINALIZE
h) Checking for the restart conditions (at each node separately), algorithm has to be restarted if it hasn’t improved the best results during several last steps
i) if the restart conditions are reached then
   - e1) if I am not the master node, start sending the best $F(X)$ value
   - e2) reset probability vector
j) go to step b
The algorithm was re-written (Figure 2) so that all the data transfer was performed at the step of checking of the restart conditions (text in bold used for the steps performing message passing). If the algorithm does not improve the local maximum values of the objective functions after $c_{\text{max}}$ generations, the process tries to communicate to the other processes to check if any other node has improved the previous global maximum. If so, the process receives that value and the corresponding $X$ vector values. If the local maximum value is better than the received one then we report our values as the new global maximums.

In our algorithm, we do not send $P$ vector values which are rather long. The vector $P$ for the problem with 10000 variables needs 40000 bytes to be sent. The node having achieved global optimum sends this value and the values of the vector $X_G^*$ (1250 bytes in case of 10000 Boolean variables). The values of the new probability vector are calculated as

$$p_i = \frac{(C_{\text{corr}} + (1-2C_{\text{corr}})x_{G,i}^*)P_G^\text{AVG}}{C_{\text{corr}} + (1-2C_{\text{corr}})P_G^\text{AVG}}. \quad (18)$$

Here, $x_{G,i}^*$ is the $i$-th element of the received vector $X_G^*$, $D$ is the dimension of the problem (number of the elements of $X_G^*$), $C_{\text{corr}}$ is some small real value, $P_G^\text{AVG}$ is the average value of the probability vector generated the vector $X_G^*$. If the received message does not contain it then the node which has received the message evaluates it:

$$p_G^\text{AVG} = \frac{\sum_{i=1}^{D} x_{G,i}^*}{D}. \quad (19)$$

Taking into consideration the constraints (3) if they exist, the value of the constant $C_{\text{corr}}$ can be calculated as $0.5/V$.

The last version of the algorithm, though it is rather efficient, has very low data interchange between nodes (log shows that there was only 31-152 data transfer sessions for each node during 1 hour). So, with this intensity of data traffic, there is no need to implement any expensive network to build an efficient MPI cluster for random search problems solution.

4 Results

In case of the problem (14), the results achieved by the program working on 8 nodes separately with comparison of the best result after the last step (4 nodes, 2 processes per node, 10000 Boolean variables, 11000 constraints, non-linear objective function) this version of the algorithm achieves in half time (2 hours for the separately working nodes and 63 minutes for this version of the algorithm).

![Figure 3. Comparison of the results of serial and parallel versions of the algorithm](image)

Traveling salesman problem with 1000 points (1000000 Boolean variables) after 4 hours of calculation on 4 nodes had not given the exact results, the results were still improving at the last steps, problem with 500 points (250000 variables) after 4 hours of calculation gave rather good result, very close to the optimal solution.

The parallel efficiency for the cluster system of 6 nodes is 0.81. An example (Figure 3) shows the results of
running of the algorithm on cluster and on a single computer which is a part of that cluster. To build this diagram, we included in our algorithm special block which stores the maximum objective function value reached by the algorithm and the current time after each 10 steps to a special array. The horizontal line shows when the algorithm achieves the control value (65015.17). To reach this control value, the serial version of the algorithm has spent 299 minutes (17748 seconds), the parallel version has spent 59 minutes (3556 seconds).

The average value of the parallel efficiency (0.81) is calculated as the average speed-up coefficient after 10 runs for 5 different objective functions. It is interesting, that at the very first steps, the parallel efficiency coefficient is less than at the following steps. The possible reason of this fact is the process of parameters tuning which is performed by each node at the first steps and the intensive message passing.

5 Conclusions

- Blocking MPI operations give negative results in comparison with non-blocking ones for the random search problems. In case of fine granularity parallelization [12, 13], results are negative even in comparison with a single node;
- Though the realization of the simplest protocol with 6 kinds of messages is rather a complicated problem, non-blocking operations work fine, their implementation gives significant (2x) speedup in comparison with the simultaneous running of the same code at separate nodes;
- Using the high-performance clusters increases the maximum scale of the problems solved;
- To build the MPI cluster for random search problems, expensive high-performance network is not needed because of very low intensity of data traffic;
- Implementation of MPI does not decrease the productivity of a single node, OpenMP still gives perfect results in comparison with serial code.

6 References

[12] Salvatore Rinaudoy, Francesco Moschellay, Marcello A. Anilez "Controlled Random Search Parallel Algorithm for Global Optimization with Distributed Processes on Multivendor CPUs"; http://citeseer.ist.psu.edu/710446.html