A Dynamic Method to Solve the Fixed Charge Network Flow Problem

Zhibin Nie * Shuning Wang **

* National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, China. (e-mail: nzb15@mails.tsinghua.edu.cn).
** National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, China. (e-mail: swang@mail.tsinghua.edu.cn)

Abstract: This paper studies the widely applied fixed charge network flow problem (FCNFP), which is NP-hard. We approximate the FCNFP with a bilinear programming problem that is determined by a parameter $\varepsilon$. When $\varepsilon$ is small enough, the optimal solution to the bilinear programming problem is the same as the optimal solution to the FCNFP. Therefore, solving the FCNFP can be transformed into solving a series of bilinear programming problems with decreasing $\varepsilon$. In this paper, these bilinear programming problems are solved by alternately solving two coupled linear programming problems. A dynamic method is proposed to update $\varepsilon$ after solving one of the linear programming problems rather than solving the whole bilinear programming problem. Numerical experiments show the performance of the proposed method.

Keywords: Network flow problem, Fixed charge, Bilinear programming, Dynamic method.

1. INTRODUCTION

The fixed charge network flow problem (FCNFP) is widely applied in the field of network optimisation, in areas such as network design (Paraskevopoulous et al. (2016)), production planning (Nasiri et al. (2014)), inventory management (Hovav and Tsadikovich (2015)) and transportation science (Moghaddam et al. (2019)). Theoretically, the FCNFP minimises the concave total cost function under linear constraints and is therefore NP-hard (Guisewite and Pardalos (1990)).

By adding a 0-1 variable for each arc to indicate whether there is some flow passing through, the FCNFP can be modelled as a mixed-integer linear programming (MILP) problem and then be solved exactly by branch-and-bound algorithms (F. Ortega (2003); Kowalski et al. (2014); Fontes et al. (2006); Palekar et al. (1990); G. Bernard (2014)). Another exact method is the vertex ranking algorithm (Murty (1968)), which is based on the property that the optimal solution to the concave minimisation problem on a convex polyhedron can be found at one of the vertices of the feasible region. However, these exact algorithms are only applicable to small-scale FCNFPs because their computational complexity increases exponentially with the problem scale. For large-scale FCNFPs, the calculations become unacceptably large.

A huge variety of heuristic algorithms have been presented to search for a suboptimal solution for large-scale FCNFPs. Walker (1976) proposed the heuristic adjacent extreme point algorithm, which escapes a local optimum by jumping over adjacent extreme points to resume iterating two or three extreme points away. Kim et al. (2006) and Kim and Pardalos (1999) proposed a dynamic slope scaling procedure (DSSP) to approximately solve the FCNFP. The DSSP transforms the FCNFP into a linear programming problem that is updated dynamically as the algorithm progresses. Nahapetyan and Pardalos (2008) transformed the FCNFP into a series of bilinear programming problems and then solved them by the adaptive dynamic cost updating procedure (ADCU). Rebennack et al. (2009) combined the ADCUP with the branch-and-bound method. They used the ADCUP to find a suboptimal solution and then employed the suboptimal solution as a starting point to solve the FCNFP exactly by the branch-and-bound algorithm. Other classical heuristic methods (Lotfi and Moghaddam (2013); Sherbiny and Alhamali (2013); Fontes and Goncalves (2007); Adlakha and Kowalski (2010); Hewitt et al. (2010)), such as the genetic algorithm and the particle swarm algorithm, were also applied to solve the FCNFP.

In these heuristic algorithms, the ADCUP has been proven to be very efficient. The objective function of the FCNFP is $f = \sum_{a \in A} f_a$, where $A$ is the set of arcs and $f_a$ is the cost function for arc $a$. In the ADCUP, the original cost function $f_a$ with a fixed charge is underestimated by a piecewise linear function $\phi_a^{\varepsilon_a}$ with parameter $\varepsilon_a$. The smaller $\varepsilon_a$ is, the smaller the difference between $\phi_a^{\varepsilon_a}$ and $f_a$. Let $\varepsilon$ be the parameter vector consisting of all elements in set $\{\varepsilon_a | a \in A\}$. Then, the FCNFP can be approximated as a bilinear programming problem with parameter $\varepsilon$. Moreover, when $\varepsilon$ is small enough, the FCNFP is equivalent to the bilinear programming problem.
problem. The ADCUP is an iterative algorithm. In each iteration, the ADCUP applies the dynamic cost updating procedure (DCUP) to find a locally optimal solution of the bilinear programming problem and then reduces the parameter \( \varepsilon \) based on the obtained solution until \( \varepsilon \) is small enough. This means that the ADCUP needs to implement a complete DCUP for each fixed \( \varepsilon \), which is not efficient enough. Structurally, the DCUP consists of solving a series of coupled linear programming problems. To accelerate the update of \( \varepsilon \), we propose the dynamic problem-updating procedure (DPUP), which updates \( \varepsilon \) after solving one linear programming problem instead of implementing the complete DCUP. In other words, the DPUP updates \( \varepsilon \) much more efficiently than the ADCUP. The quality of the solution obtained by the DPUP is also verified experimentally.

This paper is organised as follows: Section 2 formulates the FCNFP and approximates it as a bilinear programming problem. Section 3 presents the continuous bilinear algorithm and the DPUP. The validity and convergence of the DPUP are also analysed. Section 4 verifies the performance of the DPUP by numerical experiments. Section 5 concludes the paper.

2. PRELIMINARIES

2.1 The FCNFP

Let \( G = (N, A) \) be a directed network with \( n \) nodes and \( m \) arcs, where \( N \) is the set of nodes and \( A \) is the set of arcs. Each arc \( a \in A \) is associated with a flow \( x_a \), a capacity \( u_a \), and a cost function \( f_a(x_a) \). In the FCNFP, \( f_a(x_a) \) consists of two parts, the fixed cost \( s_a \) and the variable cost \( c_a x_a \), where \( c_a \) is the unit cost. Therefore, \( f_a(x_a) \) is discontinuous and can be expressed as

\[
f_a(x_a) = \begin{cases} 0 & x_a = 0, \\ s_a + c_a x_a & x_a \in (0, u_a]. \end{cases}
\]  

Let \( x \) be the flow vector, \( B \) be the node-arc incidence matrix of \( G \), and \( b \) be the node supply vector. The FCNFP can be formulated as the following problem:

**FCNFP:**

\[
\min_x f(x) = \sum_{a \in A} f_a(x_a)
\]

\[\text{s.t. } Bx = b \]

\[0 \leq x \leq u,
\]

where \( x, u \in \mathbb{R}^m \), \( b \in \mathbb{R}^n \), and \( B \in \mathbb{R}^{n \times m} \).

2.2 Approximation to the FCNFP

To address the discontinuity of \( f_a(x_a) \), we approximate \( f_a(x_a) \) as a continuous concave piecewise linear function

\[
\phi_a^\varepsilon(x_a) = \begin{cases} \varepsilon c_a x_a & x_a \in [0, \varepsilon a], \\ s_a + c_a x_a & x_a \in [\varepsilon a, u_a], \end{cases}
\]  

where \( \varepsilon a = c_a + s_a / \varepsilon \).

Plots of \( f_a(x_a) \) and \( \phi_a^\varepsilon(x_a) \) are shown in Fig. 1. We can see that \( \phi_a^\varepsilon(x_a) \) is an underestimate of \( f_a(x_a) \). Specifically, we have \( \phi_a^\varepsilon(x_a) < f_a(x_a) \) for \( x_a \in (0, \varepsilon a) \) and \( \phi_a^\varepsilon(x_a) = f_a(x_a) \) for \( x_a \in 0 \cup [\varepsilon a, u_a] \).

For any \( \varepsilon \) with \( \varepsilon a \in (0, u_a), \forall a \in A \), a continuous piecewise linear network flow problem (CPLNFP) is defined as follows:

**CPLNFP(\( \varepsilon \))**:

\[
\min_{x,y} \varphi^\varepsilon(x) = \sum_{a \in A} (\varepsilon c_a x_a + (s_a - \varepsilon a x_a) y_a)
\]

\[\text{s.t. } Bx = b \]

\[0 \leq x \leq u \]

\[0 \leq y \leq 1.
\]

By replacing the binary constraint (5) with \( 0 \leq y_a \leq 1 \), CPLNFP(\( \varepsilon \)) can be relaxed as the following continuous bilinear network flow problem (CBLNFP).

**CBLNFP(\( \varepsilon \))**:

\[
\min_{x,y} \varphi^\varepsilon(x) = \sum_{a \in A} (\varepsilon c_a x_a + (s_a - \varepsilon a x_a) y_a)
\]

\[\text{s.t. } Bx = b \]

\[0 \leq x \leq u \]

\[0 \leq y \leq 1.
\]
3. ALGORITHM

3.1 Continuous bilinear algorithm

Based on sections 2.2 and 2.3, we can solve the FCNFP by solving CBLNFP(ε) with a sufficiently small ε. However, it is difficult to obtain the applicable ε by definition. The following theorem provides a feasible method to obtain an applicable ε.

**Theorem 2.** (Rebennack et al. (2009)) For a specified ε, let (x*, y*) be the optimal solution to CBLNFP(ε). If x* ∈ 0 ∪ [εa, uA], ∀a ∈ A, (7) then x* is also the optimal solution to the FCNFP.

**Proof.** See the proof of Corollary 3.2 in Rebennack et al. (2009).

We can start from a large parameter ε and iterate to reduce ε until the optimal solution to CBLNFP(ε) satisfies condition (7). The solution obtained in the last iteration is used to update the parameter ε in the next iteration. This process can be shown as the following continuous bilinear algorithm (CBA):

**Algorithm 1** Continuous bilinear algorithm.

**Require:** Matrix B, vector b, u, parameter α ∈ (0, 1);
**Ensure:** Parameter ε, solution xε;

1: **Step 1:** Initialise: ε = u;
2: **Step 2:** Solve CBLNFP(ε) and obtain the optimal solution xε.
3: **Step 3:**
4: if ∃xε ∈ (0, ε], a ∈ A then
   - Aε = {a|a ∈ A, xε ∈ (0, ε]};
   - εa = α · εa, ∀a ∈ Aε;
   - Go to Step 2;
5: end if
6: return ε, xε;

3.2 The dynamic method to update ε

CBLNFP(ε) is still concave and thus is computationally expensive to solve. From Step 2 of **Algorithm 1**, we can see that xε is only used to update ε before the last iteration. If we can use much less time to obtain a suboptimal solution to CBLNFP(ε), which can also update ε, the efficiency of the CBA can be greatly improved.

An efficient method to obtain a suboptimal solution to CBLNFP(ε) is the variable rotation method, which alternately fixes x and y and then solves the resulting linear programming problem. The variable rotation method divides CBLNFP(ε) into the following two coupled linear programming problems, which we refer to as LPε(y) and LPε(x).

**Algorithm 2** Dynamic problem-updating procedure

**Require:** Matrix B, vector b, u, yε, parameter α ∈ (0, 1);
**Ensure:** Parameter ε, solution xε;

1: **Step 1:** Initialise:
   - ε = u;
   - yε = y0;
2: **Step 2:** Solve LPε(yε) and obtain the optimal solution xε.
3: **Step 3:**
4: if ∃xε ∈ (0, ε], a ∈ A then
   - Aε = {a|a ∈ A, xε ∈ (0, ε]};
   - εa = α · εa, ∀a ∈ Aε;
   - Solve LPε(xε) and obtain the optimal solution yε;
   - Go to Step 2;
5: end if
6: return ε, xε;

3.3 Analysis of the convergence

The following corollary shows the convergence of the DPUP for solving the FCNFP.

**Corollary 1.** The DPUP solves the FCNFP in a finite number of iterations.

**Proof.** In the worst case, each iteration of the DPUP updates the parameter εa for only one arc a ∈ A. Hence, for each arc a ∈ A, the maximum number of iterations needed is given by In = I + 1, where I is the smallest integer satisfying αIuA ≤ δ and 1 is the step needed to check the stopping criterion. In actual implementation, all the arcs in A need only one step in total to check the stopping criterion. Therefore, the upper bound on the total number of iterations of the DPUP is given by

\[
N_u = 1 + \sum_{a \in A} \max\{\lceil log_\alpha \frac{\delta}{uA} \rceil, 0\}. \tag{10}
\]

4. NUMERICAL EXPERIMENTS

Numerical experiments are conducted in MATLAB 2014a on a Windows 10 platform with an Intel Core i7 3.2 GHz...
processor and 16.0 GB of RAM. We compare the DPUP with the ADCUP and the CPLEX MIP Solver.

The ADCUP uses the complete DCUP to obtain a local optimum for CBLNFP(ε) and then updates the parameter ε. The CPLEX MIP Solver solves the following 0-1 mixed-integer programming problem, which is equivalent to the FCNFP.

\[
\text{MIP-FCNFP} : \min_{x,y} f(x) = \sum_{a \in A} c_a x_a + s_a y_a \\
\text{s.t.} \quad Bx = b \\
0 \leq x_a \leq u_a y_a, \forall a \in A \\
y_a \in \{0,1\}, \forall a \in A.
\]

(11)

For small-scale problems, the CPLEX MIP Solver can obtain the exact solution of the FCNFP. However, for large-scale problems, the CPLEX MIP Solver cannot find the optimal solution in an acceptable time. This paper uses the CPLEX MIP Solver with version 12.6 and sets the acceptable time to 200 seconds. If the CPU running time exceeds 200 seconds, we say the MIP-FCNFP cannot be solved by the CPLEX MIP Solver.

The parameter α also affects the convergence speed of the DPUP and ADPUP. For fairness in the comparison, α is set to 0.5 for both the DPUP and ADPUP.

### 4.1 Test problems

The test problems are divided into 12 problem sets according to the network scale (the number of nodes (n) and the number of arcs (m)) shown in Table 1. Each problem set consists of 10 test problems with the same network scale, where networks are randomly generated by the benchmark network generator NETGEN [16]. In each test problem, the fixed charge \( s_a \) and the unit cost \( c_a \) for any arc \( a \in A \) are randomly generated in \([50, 100]\) and \([5, 15]\), respectively. The number of supply (demand) nodes is generated uniformly between 20% and 30% of the total nodes. The total supply flow is generated uniformly between 40 and 50 times the number of nodes and then randomly assigned to the supply nodes.

For ease of expression, a solver set \( S \) is defined as

\[
S = \{D, A, C\},
\]

where “D” denotes the DPUP, “A” denotes the ADCUP and “C” denotes the CPLEX MIP Solver. When a test problem is solved by \( \zeta \in S \), we use \( f_\zeta \) and \( T_\zeta \) to denote the obtained objective function value and the CPU running time, respectively.

### 4.2 Computation results for small-scale problems

The small-scale problems in sets 1-2 are solved by the DPUP, ADCUP and CPLEX MIP Solver. The CPU running time, measured in seconds, is used to evaluate the solving efficiency. The objective function values are used to evaluate the solving accuracy. The smaller the objective function value is, the better the corresponding solution is.

Moreover, as the problem scale increases, the gap between \( T_D < T_A < T_C \) becomes wider. For the relative error, we can see that there are 13 (65%) test problems with \( \text{RE}_D < \text{RE}_A \) and 7 (35%) test problems with \( \text{RE}_D > \text{RE}_A \). That is, compared with the ADCUP, the DPUP has a higher probability of finding a better solution.

### 4.3 Computation results for large-scale problems

For the large-scale test problems in sets 3-12, the CPLEX MIP Solver cannot obtain the exact solutions in an acceptable CPU running time (200 seconds). Therefore, the test problems in these 10 sets are solved by the DPUP and ADCUP. For each test problem, the time ratio (TR) is defined as

\[ TR = T_A/T_D. \]

(14)

The average, minimum and maximum time ratios for test problems in each problem set are used to evaluate the solving efficiency.

For each test problem, since the exact solution cannot be obtained, we define the pseudo-error (PE) as

\[
\text{PE}_D(\%) = \frac{f_D - \min\{f_D, f_A\}}{\min\{f_D, f_A\}} \times 100\%,
\]

\[
\text{PE}_A(\%) = \frac{f_A - \min\{f_D, f_A\}}{\min\{f_D, f_A\}} \times 100\%.
\]

(15)

Similarly, the average and maximum pseudo-error for the test problems in each problem set are used to evaluate the solving accuracy.

Moreover, for each problem set, we define the percentage of better values (PV) to compare the overall accuracy of the solutions obtained by the DPUP and ADCUP.

\[
\text{PV}_D(\%) = \frac{N_{f_D < f_A}}{N_p} \times 100\%,
\]

\[
\text{PV}_A(\%) = \frac{N_{f_A < f_D}}{N_p} \times 100\%.
\]

(16)
Table 2. Performance of the DPUP, ADCUP and CPLEX MIP Solver for small-scale problems.

<table>
<thead>
<tr>
<th>Set</th>
<th>Problem No.</th>
<th>Size</th>
<th>CPU Running Time (seconds)</th>
<th>Objective Function Value</th>
<th>RE(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n m</td>
<td></td>
<td>TD</td>
<td>TA</td>
<td>TC</td>
</tr>
<tr>
<td>1</td>
<td>1 2 100</td>
<td>0.0032</td>
<td>0.0054</td>
<td>0.1389</td>
<td>2329</td>
</tr>
<tr>
<td>2</td>
<td>1 2000</td>
<td>0.0013</td>
<td>0.0313</td>
<td>0.1196</td>
<td>1639</td>
</tr>
<tr>
<td>3</td>
<td>1 1000</td>
<td>0.0044</td>
<td>0.0199</td>
<td>0.0721</td>
<td>2000</td>
</tr>
<tr>
<td>4</td>
<td>1 5000</td>
<td>0.0025</td>
<td>0.0301</td>
<td>0.1756</td>
<td>1647</td>
</tr>
<tr>
<td>5</td>
<td>1 7000</td>
<td>0.0021</td>
<td>0.0185</td>
<td>0.2357</td>
<td>1645</td>
</tr>
<tr>
<td>6</td>
<td>1 5000</td>
<td>0.0016</td>
<td>0.0259</td>
<td>0.0962</td>
<td>1649</td>
</tr>
<tr>
<td>7</td>
<td>1 7000</td>
<td>0.0015</td>
<td>0.0234</td>
<td>0.0858</td>
<td>1932</td>
</tr>
<tr>
<td>8</td>
<td>1 1000</td>
<td>0.0024</td>
<td>0.0416</td>
<td>0.1239</td>
<td>1636</td>
</tr>
<tr>
<td>9</td>
<td>1 2000</td>
<td>0.0019</td>
<td>0.0433</td>
<td>0.2362</td>
<td>1483</td>
</tr>
<tr>
<td>10</td>
<td>1 3000</td>
<td>0.0012</td>
<td>0.0421</td>
<td>0.2293</td>
<td>1655</td>
</tr>
</tbody>
</table>

Table 3. Performance of the DPUP and ADCUP for large-scale problems.

<table>
<thead>
<tr>
<th>Set</th>
<th>Size</th>
<th>CPU Running Time (seconds)</th>
<th>TR</th>
<th>PE (%)</th>
<th>PV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n m</td>
<td>Aver (min,max)</td>
<td>Aver (min,max)</td>
<td>Aver max</td>
<td>Aver max</td>
</tr>
<tr>
<td>3</td>
<td>120 1500</td>
<td>(0.0142,0.0335)</td>
<td>(0.0782,1.0270)</td>
<td>41</td>
<td>0.39</td>
</tr>
<tr>
<td>4</td>
<td>140 2000</td>
<td>(0.0266)</td>
<td>(1.8749)</td>
<td>62</td>
<td>0.26</td>
</tr>
<tr>
<td>5</td>
<td>160 2500</td>
<td>(0.0314)</td>
<td>(2.7123)</td>
<td>86</td>
<td>0.59</td>
</tr>
<tr>
<td>6</td>
<td>180 3000</td>
<td>(0.0894)</td>
<td>(10.136)</td>
<td>113</td>
<td>0.50</td>
</tr>
<tr>
<td>7</td>
<td>200 4000</td>
<td>(0.1071)</td>
<td>(14.4167)</td>
<td>134</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>220 5000</td>
<td>(0.1104)</td>
<td>(19.563)</td>
<td>177</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>240 6000</td>
<td>(0.1201)</td>
<td>(27.101)</td>
<td>209</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>260 7000</td>
<td>(0.1498)</td>
<td>(34.967)</td>
<td>233</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>280 8000</td>
<td>(0.1901)</td>
<td>(48.376)</td>
<td>254</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>300 9000</td>
<td>(0.2185)</td>
<td>(67.826)</td>
<td>260</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ N_{fp} < f_A \] is the number of test problems where the objective function values obtained by the DPUP are smaller than those obtained by the ADCUP. \[ N_{fp} < f_D \] is the number of test problems where the objective function values obtained by the ADCUP are smaller than those obtained by the DPUP, and \( N_p \) is the number of test problems in each problem set. In this paper, \( N_p = 10 \).

Computation results are shown in Table 3. We can see that for any problem set, the average (maximum or minimum) of \( T_D \) is always much less than the average (maximum or minimum) of \( T_A \). The average, maximum and minimum of the TR are all much greater than 1. To further demonstrate the trend of TR with the problem scale, the computation results for TR are shown in Fig. 2, where the thick line represents the average of TR and each thin vertical line represents the range between the minimum and maximum of TR. As the problem scale increases, the average of TR increases approximately linearly.

Fig. 2. Plots of the average, minimum and maximum of TR.
For each problem set, \( PE_D \) is clearly smaller than \( PE_A \), and \( PV_D \) is clearly larger than \( PV_A \). Moreover, for problem sets 7-12, \( PE_D \) is 0 and \( PV_D \) is 100%, which means that the DPUP can obtain better solutions than the ADCUP for all test problems.

4.4 Summary

Considering the solving efficiency, the DPUP always shows obvious superiority over the ADCUP in problems of different scales.

In regard to optimising capacity, neither the DPUP nor the ADCUP can obtain the global optimal solution in most cases. For small-scale problems, the accuracy of the solution obtained by the DPUP and ADCUP is related to the test problem itself. However, the DPUP statistically obtains better solutions in more test problems. For very large-scale problems, the DPUP can always obtain better solutions.

In general, compared with the ADCUP, the DPUP possesses an obvious superiority in both solving efficiency and optimising capability.

5. CONCLUSION

The motivation for our study comes from the wide application of the FCNFP. We transform the task of solving the FCNFP into solving a series of bilinear programming problems. The major contribution of this paper is the DPUP, which is used to update the bilinear programming problem dynamically. The superiority of the DPUP is that it only needs to solve a linear programming problem instead of solving the complete bilinear programming problem before updating the parameter \( \epsilon \). In numerical experiments, the performance of the DPUP is evaluated by comparison with the ADCUP and CPLEX MIP Solver in randomly generated test problems.

REFERENCES


