Partial penalty method for flow optimization in wireless networks

I.V. KONNOV
Kazan University
Dep. of System Analysis and
Information Technologies
ul. Kremlevskaya, 18, Kazan 420008
RUSSIA
konn-igor@ya.ru

O.A KASHINA
Kazan University
Dep. of Data Analysis
and Operations Research
ul. Kremlevskaya, 18, Kazan 420008
RUSSIA
olga.kashina@mail.ru

E. LAITINEN
University of Oulu
Dep. of Mathematical Sciences
POB 3000, 90014 University of Oulu
FINLAND
erkki.laitinen@oulu.fi

Abstract: We consider a general problem of optimal assignment of users to providers of wireless telecommunication networks, which minimizes the total expenses and has certain resource allocation restrictions. We show that it can be formulated as an extended transportation problem. Due to variability of demand and inexactness of data we suggest to solve this problem by a suitable penalty method. We consider both full and partial versions of this method and show that the latter has certain preferences. The computational experiments confirm these conclusions.

Key–Words: Resource allocation, wireless networks, assignment of users, transportation problem, partial penalty method.

1 Introduction

The current trends in development of information technologies imply radical worldwide modernization of industry and economy with ubiquitous implementation of wireless networks endowed with sensors, electronics, and software that communicate and interact with each other and with the environment in collecting, storing, exchanging, and processing data. This gives actually new possibilities for the development of artificial intelligence, robotics, 3D printing, nanotechnology, biotechnology, quantum computing and other breakthrough technologies.

At the same time, increasing and variable demand of information services and users movement lead to serious congestion effects, whereas significant network resources may be utilized inefficiently, especially in the case when fixed allocation mechanisms are implemented. This situation forces us to apply more flexible and dynamical allocation mechanisms; see e.g. [1, 2]. For this reason, it seems more suitable to find an approximate solution of a proper resource allocation problem, which does not require high accuracy, within an acceptable time interval rather than to calculate the exact one. Usually, the resource allocation problems are based on the utility maximization approach; see e.g. [3, 4, 5]. This utility can be treated as willingness to pay for the utilized resource.

In the previous papers, we considered the models of allocation of resource of one provider among users in a network divided into several zones. They are reduced to optimization problems with one joint and many simple constraints and admit efficient decomposition methods; see e.g. [6, 7]. However, users now can change easily telecommunication network providers for attaining better values of prices and quality of transmission, hence the more general problem of assignment of connections is very actual as well. In this paper, we just consider a general problem of optimal assignment of users to providers of wireless telecommunication networks, which minimizes the total expenses and has certain resource allocation restrictions. That is, providers have different coverage areas with the required level of service quality for each connection within such an area, whereas users have lower bounds for their volume of the resource and their desired prices. We should also take into account expenses of providers for maintaining the required volume of service. We show that the problem allows the statement in the form of the transportation problem (TP for short) with bilateral constraints on variables. We propose a technique that implies the use of penalty functions but only for certain constraints, whereas the rest constraints form a set of points having a special structure. It is used as a feasible set for an auxiliary problem. The key moment is that in spite of the presence of binding constraints, the suggested auxiliary problem is solvable by a simple finite algorithm. We have performed extensive numerical experiments that confirmed the advantage of the proposed method in comparison with the custom one involving penalization of all the constraints.

2 The problem formulation

Within certain fixed planning time period, we consider a region (territory) where wireless network services of several providers are used by mobile devices owners. Each of these users can be either a transmitter or a receiver of a signal. Denote by m the number of providers; let us numerate providers using the index i ($i = 1, \ldots, m$). Within the given time period there arise connections (signal transmissions) between certain users. Denote by n the number of (pair) connections; let us numerate connections using the index j ($j = 1, \ldots, n$). Signal transmissions require certain expenditures of providers' resources (say, the bandwidth or power of the wireless channel). It is natural to assume that the resource amount possessed by each provider i is bounded by some value γ_i. Let the symbol $x_{i,j}$ stand for the unknown amount of the resource allotted by provider i for pair connection j (below for brevity we just say "flow (i, j)"). Denote by $\alpha_{i,j}$ the upper bound for flow (i, j) and by β_j the lower bound for the total flow for connection j. Let b_j be the price (willingness to pay) proposed by pair j and let $a_{i,j}$ be expenses per unit for connection j incurred by provider i. Then the pure total expenses are given by the expression

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}x_{i,j} - \sum_{j=1}^{n} b_j \left(\sum_{i=1}^{m} x_{i,j}\right) \equiv \sum_{i=1}^{m}\sum_{j=1}^{n} c_{i,j}x_{i,j},$$

where $c_{i,j} \equiv a_{i,j} - b_j$. The goal is to minimize the pure total expenses due to a proper distribution of the load upon network providers.

Note that any connection can be supported and accomplished at the proper service level only by selected providers in accordance with their quality service coverage areas. That is, each connection j can be accomplished by selected providers whose indices belong to the set P_j. However, for all $i \notin P_j$ we can set $\alpha_{i,j} = 0$, which implies $x_{i,j} = 0$. Therefore, without loss of generality we can consider only the case where $P_j = \{1, \ldots, m\}$ for each $j \in \{1, \ldots, n\}$. The problem takes the form

$$\min \longrightarrow \sum_{i=1}^{m}\sum_{j=1}^{n} c_{i,j}x_{i,j}, \tag{1}$$

subject to

$$\sum_{i=1}^{m} x_{i,j} \geq \beta_j, \quad j = 1, \ldots, n, \tag{2}$$

$$\sum_{j=1}^{n} x_{i,j} \leq \gamma_i, \quad i = 1, \ldots, m, \tag{3}$$

$$0 \leq x_{i,j} \leq \alpha_{i,j},\ i = 1, \ldots, m,\ j = 1, \ldots, n. \tag{4}$$

Problem (1)–(4) is nothing but the so-called open transportation problem with bilateral constraints on variables. It becomes the classical transportation problem in the case when $\alpha_{i,j} = +\infty$ for all i, j; see [8] for more details and references.

In spite of the existence of finite solution methods for the TP (see, for example, [8]), we intend to apply some other iterative methods for this problem. The most influential factor that affects the applicability of exact methods for solving the TP, evidently, is the fast growth of the problem dimension, which, in turn, leads to the accumulation of computation errors and poor conditionality of the constraint coefficient matrix. Moreover, in practice, the feasible set of the open TP is not necessarily nonempty. In such cases one can find a solution close to the optimal (feasible) one only by approximate methods. Another factor that contributes to the relevance of the development of approximate solution methods for the TP is the appearance of new applications of the transportation model; for example, along with classical applications in the optimization of production, transportation, and sales of some commodity, this model appears to be applicable in the optimization of the performance of mobile networks. Such a problem usually has a large dimensionality, and its initial data are inexact and nonstationary. Moreover, in practice, problem (1)–(4) are often being solved in order to estimate certain characteristics of the network performance; in this case it is more important to find an appropriate solution of the problem within an acceptable time frame rather than to obtain a high accuracy solution.

In this paper we propose an approximate solution method for problem (1)–(4) which is based on application of penalty functions.

3 The partial penalty method

As distinct from the custom penalty method, in the partial penalty method (PPM for short) we impose penalties only on selected constraints. The set formed by the rest constraints has a special structure which allows us to solve the corresponding auxiliary problem by a simple finite algorithm. Thus we intend to attain higher quality of solutions.

First we introduce the so-called cut function

$$[t]_+ = \max\{0, t\},$$

and then define the penalty function for the constraints in (3):

$$\Phi(X) \equiv \sum_{i=1}^{m} \left[\sum_{j=1}^{n} x_{i,j} - \gamma_i\right]_+^2. \tag{5}$$

We take a positive penalty parameter τ and define the auxiliary function

$$\Psi(X, \tau) = \langle C, X \rangle + \tau \Phi(X). \tag{6}$$

Hereinafter C and X are $m \times n$-matrices and the denotation $\langle C, X \rangle$ stands for the double sum

$$\langle C, X \rangle \equiv \sum_{i=1}^{m} \sum_{j=1}^{n} c_{i,j} x_{i,j}. \tag{7}$$

We treat the matrix X as a point (in the space of $m \times n$-matrices). Denote the sets of points satisfying the inequalities in (4) and (2) by A and B, respectively, and set

$$X^*(\tau) \equiv \arg \min_{X \in A \cap B} \Psi(X, \tau). \tag{8}$$

Note that the function in (6) is continuous by definition and the set $A \cap B$ is closed and bounded, hence the point in (8) exists for any τ. Let us construct an iteration sequence $\{X_k^*(\tau_k)\}$, where k is the iteration number, such that the sequence $\{\tau_k\}$ is positive, increasing, and tending to $+\infty$ as $k \to \infty$, while each point $X_k^*(\tau_k)$ obeys formula (8) with $\tau = \tau_k$. Since the set $A \cap B$ is bounded, so is the sequence $\{X_k^*(\tau_k)\}$, which means that it has limit points as $k \to \infty$ and all these limit points X^* are solutions of problem (1) – (4) (see, for example, [9], Section 7.1). Moreover, this is the case for some approximations of points $x_k^*(\tau_k)$, $k = 0, 1, \dots$ Let us now consider the technique for finding the points $X_k^*(\tau_k)$, $k = 0, 1, \dots$

4 Solution of the auxiliary problems

Assume that certain real numbers $d_{i,j}$, $i = 1, \dots, m$; $j = 1, \dots, n$, are given (we concretize them below). Denote the corresponding $m \times n$-matrix by D. Let us use the denotation $\langle D, X \rangle$ in the sense of formula (7) with the symbol D in place of C. Let us describe an algorithm which solves the problem

$$\min_{X \in A \cap B} \to \langle D, X \rangle. \tag{9}$$

Let us show that in spite of the existence of constraints (2) which bound the problem variables, problem (9) falls into n independent problems which are solvable explicitly. Fix some connection $p \in \{1, \dots n\}$ and describe the algorithm for finding components $x_{i,p}$, $i = 1, \dots, m$, of a solution x to problem (9). Since this algorithm solves the auxiliary problem, we call it "Algorithm A", for short.

Algorithm A. Step 0. Given p, number providers in ascending order of $d_{i,p}$ and thus get a set of numbers $I \equiv \{i_1, \dots, i_m\}$. Introduce a new variable s and put $s := 1$.

Step 1. If

$$\sum_{i=i_1}^{i_s} \alpha_{i,p} < \beta_p,$$

then put $x_{i_s,p} := \alpha_{i_s,p}$ and go to Step 2; otherwise put

$$x_{i_s,p} := \beta_p - \sum_{i=i_1}^{i_{s-1}} \alpha_{i,p},$$

do $x_{i_v,p} := 0$ for $v = s+1, \dots, m$, and Algorithm A stops.

Step 2. If $s < m$, then put $s := s + 1$ and go to Step 1; otherwise Algorithm A stops.

Evidently, sequentially applying Algorithm A for $p = 1, \dots, n$, in n steps we get a point $\tilde{X}(D)$, whose feasibility and optimality for problem (9) is evident, provided that $A \cap B \neq \emptyset$ (in what follows we assume that this condition is fulfilled).

Let us now consider the basic problem

$$\min_{X \in A \cap B} \to \Psi(X, \tau) \tag{10}$$

for finding a point satisfying (8) with some fixed $\tau > 0$. We can solve problem (10) by the well-known conditional gradient method (CGM for short) (see, for example, [10]). Let us fix arbitrary indices $i_0 \in \{1, \dots, m\}$, $j_0 \in \{1, \dots, n\}$, and a number $\tau > 0$ and write the partial derivative of the function in (6) at a point X with respect to the variable x_{i_0,j_0}:

$$\frac{\partial \Psi(X, \tau)}{\partial x_{i_0,j_0}} = c_{i_0,j_0} + 2\tau \left[\sum_{j=1}^{n} x_{i_0,j} - \gamma_{i_0} \right]_+. \tag{11}$$

Denote by $\Psi'(X, \tau)$ the $m \times n$-matrix composed of elements (11) and treat it as the gradient of the function $\Psi(X, \tau)$ at the point X with fixed τ. Let us now describe CGM applied to problem (10).

(CGM). Step 0. Given $\tau > 0$, choose a point $X^0 \in A \cap B$. Assume that a point X^l is known already; $l = 0, 1, \dots$ Let us describe the way to find the next point X^{l+1}.

Step 1. Find a solution Z^l to the linear programming problem

$$\min_{X \in A \cap B} \to \langle \Psi'(X^l, \tau), X \rangle, \tag{12}$$

and go to Step 2.

Step 2. Calculate

$$\lambda_l := \arg \min_{\lambda \in [0,1]} \Psi(\lambda X^l + (1 - \lambda) Z^l, \tau) \tag{13}$$

and put $X^{l+1} := \lambda_l X^l + (1 - \lambda_l) Z^l$, $l := l + 1$ and go to Step 1.

For each $l = 0, 1, \dots$ by putting $D := \Psi'(X^l, \tau)$ we get problem (9) in (12) and solve it by Algorithm A. Problem (13) can be solved by any one-dimensional minimization method (see, for example, [9], Section 3.7). In numerical experiments we used the well-known golden section method (see, for example, [9], p. 84).

5 The usual penalty method

As distinct from the PPM, where the penalty function is introduced only for constraints in (3). In the usual (or full) penalty method (FPM for short) we define penalty functions for both groups of constraints, namely, for those in (3) and (2):

$$\tilde{\Phi}(X) \equiv \sum_{i=1}^{m} \left[\sum_{j=1}^{n} x_{i,j} - \gamma_i \right]_+^2 + \sum_{j=1}^{n} \left[\beta_j - \sum_{i=1}^{m} x_{i,j} \right]_+^2,$$

and

$$\tilde{\Psi}(X, \tau) \equiv \langle C, X \rangle + \tau \tilde{\Phi}(X), \qquad (14)$$

where τ is a positive penalty parameter. We now outline the main differences from the PPM.

The auxiliary problem which is solved at each step k of the FPM consists in finding the point

$$X^*(\tau_k) \equiv \arg\min_{X \in A} \tilde{\Psi}(X, \tau_k)$$

for $k = 0, 1, \dots$ Analogously, we can solve this auxiliary problem by the conditional gradient method (CGM). Its each iteration involves a solution to the linear programming problem

$$\min_{X \in A} \rightarrow \langle \tilde{\Psi}'(X^l, \tau), X \rangle \qquad (15)$$

with $\tau = \tau_k$. The components of the gradient $\tilde{\Psi}'(X, \tau)$ in view of (14) obey the formula

$$\frac{\partial \Phi(X, \tau)}{\partial x_{i_0, j_0}} = c_{i_0, j_0} + 2\tau \left[\sum_{j=1}^{n} x_{i_0, j} - \gamma_{i_0} \right]_+ - 2\tau \left[\beta_{j_0} - \sum_{i=1}^{m} x_{i, j_0} \right]_+ .$$

Since its feasible set A represents a rectangle, problem (15) falls into $m \times n$ independent one-dimensional problems, each of them is solved explicitly. The other parts are implemented similarly.

6 Results of numerical experiments

We have numerically tested the described methods via the package Wolfram Research Mathematica 9.0.1.0 by using a computer with Processor Intel® Core^{TM} i5-430M (4M Cache, 2.26 GHz). In order to prove the efficiency of the new method (PPM) we compared the results of solving problem (1) – (4) with those of (FPM). We used the same rule for decreasing values of accuracy of inner problems. For changing the penalty parameter we used the rule $\tau_{k+1} := 2\tau_k$.

We modeled the initial data of the problem so as to know its optimum point (and, correspondingly, the exact optimal value of the objective function F^*). We stopped the process when either the absolute value of the relative deviation of the current approximation to the optimal value of the objective function from F_{opt} was not greater than 10% or the norm of the difference of neighboring points was less than some predefined value ε (we put $\varepsilon := 0.001$). For each concrete problem (i.e. concrete collection of initial data) we performed 10 tests for both methods, randomly choosing an initial point. In what follows the subscript h stands for the test number (within a series of 10 tests); symbols $F_{h(FPM)}$ and $F_{h(PPM)}$ denote, respectively, approximate values of the objective function of problem (1) – (4) calculated by FPM and PPM at test number h; symbols $\overline{F}^*_{PPM}$ and $\overline{F}^*_{PPM}$ stand, respectively, for average values of $F_{h(FPM)}$ and $F_{h(PPM)}$ in each series of 10 tests, i.e.

$$\overline{F}^*_{FPM} = \frac{\sum_{h=1}^{10} F_{h(FPM)}}{10}; \; \overline{F}^*_{PPM} = \frac{\sum_{h=1}^{10} F_{h(PPM)}}{10};$$

the relative approximation errors

$$\frac{\overline{F}^*_{FPM} - F_{opt}}{F_{opt}} \quad \text{and} \quad \frac{\overline{F}^*_{PPM} - F_{opt}}{F_{opt}};$$

and values $\bar{t}_{FPM}$ and $\bar{t}_{PPM}$ are average time consumptions in a series of 10 tests. These values are given in Table 1.

According to results shown in Table 1, with small m (not greater than 20) PPM attains the given accuracy with respect to the value of the objective function (in our tests the allowed error was 10%) much faster than FPM. Moreover, the actual error introduced by PPM has never exceeded 2.17%; mainly it was even less than 0.5%, whereas the the actual error introduced by PPM was mostly greater than 3%, sometimes approaching (or even attaing) the limit admissible value of 10%. We also calculated the ratios

Table 1: Comparison of FPM and PPM

Table 1.									
#	m	n	F_{opt}	Avg. F_{opt}		Avg. err. (%)		Avg. t (sec)	
				$\overline{F}^*_{FPM}$	$\overline{F}^*_{PPM}$	$\overline{E}_{FPM}$	$\overline{E}_{PPM}$	$\bar{t}_{FPM}$	$\bar{t}_{PPM}$
1	3	20	230.22	236.78	230.22	2.85	0.00	0.11	0.03
2	3	20	422.03	431.31	422.03	2.19	0.00	0.13	0.03
3	3	20	376.24	397.35	376.24	5.60	0.00	0.04	0.03
4	10	20	1895.37	2057.44	1902.22	8.55	0.36	0.78	0.44
5	10	20	1596.32	1614.69	1600.98	1.15	0.29	0.37	0.08
6	10	20	1159.42	1201.77	1162.29	3.65	0.24	0.13	0.11
7	10	50	2089.31	2269.17	2089.31	8.61	0.00	1.30	0.13
8	10	50	1097.54	1986.86	1940.41	4.16	1.72	0.84	0.13
9	10	50	1856.6	1944.87	1892.4	4.75	1.93	0.86	0.56
10	10	100	6108.86	6335.65	6108.86	3.71	0.00	8.55	0.26
11	10	100	6907.88	7598.67	6943.07	10.00	0.52	15.75	3.97
12	10	100	7570.59	8326.72	7678.28	9.98	1.42	1.18	0.87
13	10	1000	46905.2	50378.4	47923.77	7.40	2.17	400.38	3.72
14	10	1000	55520.4	57728.4	55520.4	3.97	0.00	61.53	4.99
15	10	1000	49232.6	51894.4	49312.4	5.41	0.16	150.96	5.86
16	2	2000	22248.8	22556.4	22573.1	1.38	1.46	9.89	2.63
17	3	2000	77472.0	79799.0	77472.0	3.00	0.00	77.20	4.78
18	3	2000	58028.2	59964.6	58028.2	3.34	0.00	479.63	4.59
19	3	2000	43151.8	44687.6	43151.8	3.56	0.00	206.12	4.66
20	3	3000	200776.0	202756.0	200776.0	0.99	0.00	376.06	10.49
Average values						4.91	0.540	89.64	2.42
21	20	20	2780.15	2780.26	2786.39	0.004	0.22	0.31	0.52
22	20	20	5111.52	5284.78	5114.05	1.43	0.05	0.20	2.46
23	20	20	5137.82	5311.33	5140.05	3.37	0.04	0.26	23.34
Average values						1.60	0.10	0.26	8.77

$$\frac{F_{max(FPM)} - F_{min(FPM)}}{\overline{F}_{(FPM)}} \quad \text{and}$$
$$\frac{F_{max(PPM)} - F_{min(PPM)}}{\overline{F}_{(PPM)}}$$

(after performing a series of 10 tests) in order to study the sensitivity of these methods to the choice of the initial point.

As appeared, both methods are insensitive to the choice of an initial point (not necessarily a feasible one), since these characteristics always equaled zero. It is evident that PPM gives better results both with respect to time and to the solution accuracy (which was much less than the allowed value of 10%). As expected, the advantage of PPM over FPM was more evident when m is small (not greater than 3) and n is very large (up to 3000), whereas the growth of m (with fixed n) impairs the performance of both methods at approximately the same rate. In certain cases time consumption of PPM was even greater than that of FPM. For example, the case when $m = 2$ and $n = 2000$ (i.e., the number of variables equals 4000) the time consumption equals 9.89 and 2.63 sec. for FPM and PPM, respectively, (see row 16 in Table 1). There were some examples with $m = 20$ and $n = 20$, where PPM showed better performance. In general, PPM appeared more efficient than FPM in most examples and is suitable for calculations. Nevertheless, due to the necessity of tuning several parameters, its convergence needs further investigations.

7 Conclusion

We considered a general problem of optimal assignment of users to providers of wireless telecommunication networks and showed that it can be formulated as an extended transportation problem. We suggested to solve this problem by a suitable penalty method instead of the exact one. We considered both full and partial versions of this method and showed that the latter had certain preferences. The computational experiments confirmed these conclusions.

Acknowledgements: In this work, the first two authors were supported by the RFBR grant, project No. 16-01-00109a. Also, the first and third authors were supported by grant No. 297689 from Academy of Finland.

References:

[1] Cheng, X., Huang, X., and Du, D.-Z., Eds., Ad Hoc Wireless Networking, Kluwer, Dordrecht, 2004.

[2] Stańczak, S., Wiczanowski, M., and Boche, H. Resource Allocation in Wireless Networks. Theory and Algorithms, Springer, Berlin, 2006.

[3] Kelly, F.P., Maulloo, A.K. and Tan, D.K.H. Rate control in communication networks: shadow prices, proportional fairness and stability, J. Oper. Res. Soc. 49, 237–252 (1998)

[4] Huang, J., Berry, R.A., and Honig, M.L. Auction-based spectrum sharing, ACM/Springer Mobile Networks and Appl. 11, 405–418 (2006)

[5] Koutsopoulos, I. and Iosifidis, G. Auction mechanisms for network resource allocation, in: Proc. of Workshop on Resource Allocation in Wireless Networks (WiOpt 2010), 554–563 (2010)

[6] Konnov, I.V., Kashina, O.A., Laitinen, E.: Optimisation problems for control of distributed resources. Int. J. Model., Ident. and Contr., 14, 65–72 (2011)

[7] Konnov, I.V., Kashina, O.A., Laitinen, E.: Two-level decomposition method for resource allocation in telecommunication network. Int. J. Dig. Inf. Wirel. Comm., 2, 150–155 (2012)

[8] Gol'shtein, E.G. and Yudin, D.B. Transportation Type Linear Programming Problems, Nauka, Moscow, 1969 [in Russian].

[9] Konnov, I.V. Nonlinear Optimization and Variational Inequalities, Kazan Univ. Press, Kazan, 2013 [in Russian].

[10] Levitin, E.S. and Polyak, B.T. Constrained minimization methods. USSR Comput. Maths. Math. Phys. 6, 1–50 (1966)