Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Service System Design with Immobile Servers, Stochastic Demand, and Congestion

Samir Elhedhli,

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 8, No. 1, Winter 2006, pp. 92–97
ISSN 1523-4614 | EISSN 1526-5498 | 06 | 0801 | 0092
DOI 10.1287/msom.1050.0094

Service System Design with Immobile Servers, Stochastic Demand, and Congestion

Samir Elhedhli
Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada,
elhedhli@engmail.uwaterloo.ca

The service system design problem seeks to locate a set of service facilities, allocate enough capacity, and assign stochastic customer demand to each of them, so as to minimize the fixed costs of opening facilities and acquiring service capacity, as well as the variable access and waiting costs. This problem is commonly known in the location literature as the facility location problem with immobile servers, stochastic demand, and congestion. It is often set up as a network of $M/M/1$ queues and modeled as a nonlinear mixed-integer program (MIP). Because of the complexity of the resulting model, the current literature focuses on approximate and/or heuristic solution methods. This paper proposes a linearization based on a simple transformation and piecewise linear approximations and an exact solution method based on cutting planes. This leads to the exact solution of models with up to 100 customers, 20 potential service facilities, and 3 capacity levels.

Key words: service system design; facility location; immobile servers; stochastic demand; congestion; nonlinear MIP; piecewise linearization; cutting plane methods
History: Received: December 19, 2003; accepted: September 27, 2005. This paper was with the authors 6 months for 3 revisions.

The most common objective in the design of service systems is to balance service capacity costs against service quality costs. The former are related to opening a service facility and allocating the necessary capacity (servers) to it, whereas the latter are often measured in terms of the probability of serving a given number of customers within a time limit, the average number of customers waiting for service, or the average waiting time per customer. From a customer's perspective, the choice of service facilities may not only be based on easy access (proximity), but also on service quality, often measured in terms of average waiting time. Thus it seems reasonable to design service systems with the objective of minimizing facility opening, capacity allocation, customer access, and customer waiting costs. This class of problems is commonly called the service system design problem (Amiri 1997). The problem environment is characterized by independent and stochastic customer demand that has to be satisfied from a set of immobile service facilities. According to Brandeau et al. (1995), this is an environment with negative congestion externalities and inelastic demand. The location literature refers to this problem as the facility location problem with immobile servers, stochastic demand, and congestion (Berman and Krass 2002). The proposed models are based on the common and well-justified assumption of Poisson arrivals and exponential service rates (Amiri 1997, 2001; Talim et al. 2001; Rajagopalan and Yu 2001).

There are two different ways in which the literature considers service quality. The first approach includes a probability constraint that ensures that waiting time or queue length does not exceed a certain threshold (see, for example, Marianov and Serra 1998, 2002). The second approach incorporates the service cost directly in the objective function. The latter approach is adopted in this paper as it explicitly balances service costs against other costs. The same is done in Amiri (1997), Wang et al. (2002), and Castillo et al. (2002). The resulting model is a mixed-integer program (MIP) that has a nonlinear objective function. To date, proposed solution methods are only heuristic or approximate, dealing with the nonlinearity of the objective function through approximations. Amiri (1997) proposes a Lagrangean heuristic, where the nonlinearity is isolated into a nonlinear knapsack subproblem whose linear programming relaxation is solved. Wang et al. (2002) propose two heuristics: (1) a greedy drop heuristic and (2) an improvement heuristic using tabu search.

Castillo et al. (2002) use an asymptotic approximation of $M/M/s$ queues to reduce the nonlinear MIP to a tractable problem.

In this paper, we overcome the nonlinearity of the model by transforming it into an equivalent linear MIP with a large, possibly exponential number of constraints through a simple transformation and piecewise linear approximations. We solve the resulting MIP by using a cutting plane method. The proposed linearization opens the door toward the use of other solution methods that can deal with a large, possibly exponential number of constraints.

The proposed methodology can easily handle side constraints and can be used within a decomposition scheme to solve larger and more complex models.

This paper is organized as follows. Section 1 describes the problem formulation and the nonlinear MIP. Section 2 details the linearization and the proposed exact solution procedure. Section 3 reports on numerical results for varying cost parameters. Section 4 concludes with the main findings and future research directions.

1. Problem Formulation

Using indices $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $k = 1, \ldots, K$ for customers, facilities, and capacity levels, respectively, we let x_{ij} and y_{jk} to be binary variables that take value 1 if customer i's demand is allocated to service facility j, and facility j is allocated capacity level k, respectively. We denote the mean demand rate of customer i by λ_i and the mean service rate of service facility j when allocated capacity level k by μ_{jk}. We additionally denote the unit access cost when customer i is assigned to service facility j by c_{ij}, the average waiting cost per unit time by t, and the fixed cost of opening a facility at site j and equipping it with capacity level k by f_{jk}. We assume that customer demand are independent Poisson distributed and are entirely assigned to a single service facility, whereas service rates are exponentially distributed. Therefore, each service facility acts as an $M/M/1$ queue with mean demand of $\lambda_j = \sum_{i=1}^{n} \lambda_i x_{ij}$ and mean service rate of $\mu_j = \sum_{k=1}^{K} \mu_{jk} y_{jk}$, for which the average waiting time is $1/(\mu_j - \lambda_j)$. The service system design problem is formulated as

$$[\text{NMIP}]:\ \min \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\lambda_i x_{ij} + t\sum_{j=1}^{n} \frac{\sum_{i=1}^{m}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk} - \sum_{i=1}^{n}\lambda_i x_{ij}} + \sum_{j=1}^{n}\sum_{k=1}^{K} f_{jk}y_{jk} \quad (1)$$

$$\text{s.t.}\ \sum_{i=1}^{m}\lambda_i x_{ij} - \sum_{k=1}^{K}\mu_{jk}y_{jk} \le 0 \quad j=1,\ldots,n \quad (2)$$

$$\sum_{k=1}^{K} y_{jk} \le 1 \quad j=1,\ldots,n \quad (3)$$

$$\sum_{j=1}^{m} x_{ij} = 1 \quad i=1,\ldots,m \quad (4)$$

$$x_{ij} \in \{0,1\} \quad i=1,\ldots,m;\ j=1,\ldots,n \quad (5)$$

$$y_{jk} \in \{0,1\} \quad j=1,\ldots,n;\ k=1,\ldots,K. \quad (6)$$

The objective function (1) minimizes access, waiting, facility opening, and capacity acquisition costs. Constraints (2)–(4) ensure steady-state conditions ($\lambda_j < \mu_j$), at most one capacity level is allocated to each facility, and each customer is assigned to exactly one service facility, respectively. The literature looks at two different cases, depending on whether x_{ij} are binary (single sourcing) or fractional (multiple sourcing).

2. A Linearization and an Exact Solution Approach

In this section, we linearize the fractional objective function of [NMIP] and propose an exact solution procedure. For this purpose, let us introduce a new nonnegative variable

$$R_j = \frac{\sum_{i=1}^{m}\lambda_i x_{ij}}{\sum_{k=1}^{K}\mu_{jk}y_{jk} - \sum_{i=1}^{m}\lambda_i x_{ij}}, \quad j=1,\ldots,n,$$

which implies that

$$\sum_{i=1}^{m}\lambda_i x_{ij} = \frac{R_j}{1+R_j}\sum_{k=1}^{K}\mu_{jk}y_{jk}.$$

Let us also introduce the variables

$$z_{jk} = \frac{R_j}{1+R_j} y_{jk}, \quad j=1,\ldots,n,\ k=1,\ldots,K.$$

We then have

$$\sum_{i=1}^{m}\lambda_i x_{ij}=\sum_{k=1}^{K}\mu_{jk}z_{jk}\quad\text{and}$$

$$z_{jk}=\begin{cases}0 & \text{if } y_{jk}=0\\ \dfrac{R_j}{1+R_j} & \text{if } y_{jk}=1.\end{cases}\tag{7}$$

Being concave, the functions $f(R_j)=R_j/(1+R_j)$, $j=1,\dots,n$ can be written as the minimum of a set of tangent piecewise linear functions. Consider nonnegative points R_j^h indexed by set H, then,

$$\frac{R_j}{1+R_j}=\min_{h\in H}\left\{\frac{1}{(1+R_j^h)^2}R_j+\left(\frac{R_j^h}{1+R_j^h}\right)^2\right\},\quad j=1,\dots,n.\tag{8}$$

Equation (8) implies that

$$\frac{R_j}{1+R_j}\le\frac{1}{(1+R_j^h)^2}R_j+\left(\frac{R_j^h}{1+R_j^h}\right)^2,\quad\forall h\in H;\ j=1,\dots,n.$$

Therefore (7) leads to

$$\begin{cases}z_{jk}=0 & & \text{if } y_{jk}=0\\ z_{jk}\le\dfrac{1}{(1+R_j^h)^2}R_j+\dfrac{(R_j^h)^2}{(1+R_j^h)^2}, & \forall h\in H & \text{if } y_{jk}=1,\end{cases}$$

which are modeled using the following set of constraints:

$$z_{jk}\le\frac{1}{(1+R_j^h)^2}R_j+\frac{(R_j^h)^2}{(1+R_j^h)^2},\quad\forall h\in H;\ j=1,\dots,n;\ k=1,\dots,K.$$

$$0\le z_{jk}\le y_{jk},\quad j=1,\dots,n;\ k=1,\dots,K.$$

Therefore, problem [NMIP] is equivalent to

[MIP(H)]:

$$\min\ \sum_{i=1}^{n}\sum_{j=1}^{m}c_{ij}\lambda_i x_{ij}+t\sum_{j=1}^{m}R_j+\sum_{j=1}^{m}\sum_{k=1}^{K}f_{jk}y_{jk}\tag{9}$$

$$\text{s.t.}\ \sum_{i=1}^{n}\lambda_i x_{ij}-\sum_{k=1}^{K}\mu_{jk}z_{jk}=0\quad j=1,\dots,n\tag{10}$$

$$\sum_{j=1}^{m}x_{ij}=1\quad i=1,\dots,m\tag{11}$$

$$\sum_{k=1}^{K}y_{jk}\le 1\quad j=1,\dots,n\tag{12}$$

$$z_{jk}-y_{jk}\le 0\quad j=1,\dots,n\tag{13}$$

$$z_{jk}-\frac{1}{(1+R_j^h)^2}R_j\le\frac{(R_j^h)^2}{(1+R_j^h)^2}\quad j=1,\dots,n;\ k=1,\dots,K;\ h\in H\tag{14}$$

$$R_j\ge 0\quad j=1,\dots,n\tag{15}$$

$$x_{ij}\in\{0,1\}\quad i=1,\dots,m;\ j=1,\dots,n\tag{16}$$

$$y_{jk}\in\{0,1\},\quad 0\le z_{jk}\le 1\quad j=1,\dots,n;\ k=1,\dots,K.\tag{17}$$

It is worth mentioning that because we minimize the objective function (9), for any facility j and capacity level k, at least one of the constraints in (14) will be binding at optimality forcing

$$z_{jk}=\min_{h\in H}\left\{\frac{1}{(1+R_j^h)^2}R_j+\left(\frac{R_j^h}{1+R_j^h}\right)^2\right\},\quad j=1,\dots,n\quad\text{when}\quad y_{jk}=1.$$

The nonlinearity of [NMIP] was eliminated at the expense of having to deal with a large number of constraints in [MIP(H)]. The large number of constraints is suitable for a cutting plane (constraint generation) approach, starting with an initial set of constraints and adding the rest as needed.

At iteration q, where a finite set of constraints (14) are generated at points $(R_j^h)_{H^q\subset H}$ indexed by set H^q, [MIP(H^q)] is a relaxation of [MIP(H)]. Hence its optimal objective, denoted by v(MIP(H^q)), provides a lower bound LB^q to [MIP(H)], and equivalently to [NMIP]. Furthermore, the solution of [MIP(H^q)], denoted by $(\bar{x}^q,\bar{y}^q,\bar{z}^q,\bar{R}^q)$, provides a feasible solution $(\bar{x}^q,\bar{y}^q)$ to [NMIP]. Hence

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda_i c_{ij}\bar{x}_{ij}^q+\sum_{j=1}^{m}\sum_{k=1}^{K}f_{jk}\bar{y}_{jk}^q+t\sum_{j=1}^{m}\frac{\sum_{i=1}^{n}\lambda_i\bar{x}_{ij}^q}{\sum_{k=1}^{K}\mu_{jk}\bar{y}_{jk}^q-\sum_{i=1}^{n}\lambda_i\bar{x}_{ij}^q}$$

is an upper bound UB^q to [NMIP]. Although the lower bound LB is monotonic, the upper bound is not. So, at iteration q, the best lower bound LB is set to LB^q, whereas the best upper bound UB is set to $\min\{UB,UB^q\}$. If the lower bound LB equals to the upper bounds UB, then $(\bar{x}^q, \bar{y}^q)$ is an optimal solution to [NMIP]. If not, a new candidate point

$$R_j^{h_{\text{new}}} = \frac{\sum_{i=1}^n \lambda_i \bar{x}_{ij}^q}{\sum_{k=1}^K \mu_{jk}\bar{y}_{jk}^q - \sum_{i=1}^n \lambda_i \bar{x}_{ij}^q}, \quad j = 1, \ldots, n$$

is used to generate further cuts in (14). The choice of the new point $R_j^{h_{\text{new}}}$ is motivated by the fact that when

$$\bar{R}_j^q = \frac{\sum_{i=1}^n \lambda_i \bar{x}_{ij}^q}{\sum_{k=1}^K \mu_{jk}\bar{y}_{jk}^q - \sum_{i=1}^n \lambda_i \bar{x}_{ij}^q} = R_j^{h_{\text{new}}}, \quad j = 1, \ldots, n,$$

the lower and upper bounds coincide and the optimal solution is found. As long as there is no cut in (14) corresponding to $R_j^{h_{\text{new}}}$, $LB \neq UB$ and optimality is not reached. Hence we add constraints (14) corresponding to point $R_j^{h_{\text{new}}}$.

The procedure is repeated until the lower and upper bounds coincide. In the appendix, we prove that at each iteration, at least one new point (and corresponding cut) is introduced, meaning that cycling cannot occur. Because (x_{ij}^q, y_{jk}^q) are binary, there is a finite set of values that $R_j^{h_{\text{new}}}$ can take. Thus the proposed method has finite convergence.

3. Numerical Testing

The proposed solution methodology is coded in Matlab 6.0 with the solution of the linear MIP problems being done in CPLEX 8.1. The tests were carried on a Sun Ultra 10/440 workstation. The test problems are based on a standard set of problems that are proposed by Holmberg et al. (1999) for the capacitated facility location problem with single sourcing. We picked three sets of problems that have a maximum of 100 customer demand points. The sets are denoted by p1–p12, p13–p24, and p41–p55.

We modify these test problems by allowing each service facility to have up to three service capacity levels $V_{jk} = kV_j$, $k = 1, 2, 3$, $j = 1, \ldots, n$. To reflect economies of scale, the fixed costs are set to $f_{jk} = u^{2/(1+k)} \times V_{jk}$ (so that $f_{j1} = u_j V_{j1}$, $f_{j2} = u_j^{2/3} V_{j2}$, $f_{j3} = u_j^{1/2} V_{j3}$), where $u_j = (f_j/V_j)$ is the unit capacity cost, and V_j and f_j are the capacity and fixed costs used in the original instances.

Table 1 displays the problem name, the number of demand points (m), the number of potential facility sites (n), the percentage of cost attributed to each component of the objective: access costs (AC), fixed costs (FC), and waiting costs (WC), the CPU time in seconds, and the number of iterations (constraints generated) that the solution method takes for different values of the waiting cost t. We set t to $\beta \times \max_{i,j} \lambda_i c_{ij}$ for $\beta = 0.1$, 1, 10, 100, and 1,000, respectively. We start with a well-placed initial set of cuts and use a stopping criteria of $(UB - LB)/UB \leq 10^{-3}$ for the algorithm.

The computational results reveal the stability and efficiency of the proposed solution method for different percentages of access, fixed, and waiting costs. The CPU times are an average 245, 290, 400, 570, and 472 seconds, while the number of constraints generated is an average of 4, 3, 4, 6, and 5 for $\beta = 0.1, 1, 10, 100$, and 1,000, and average waiting costs of 6, 15, 35, 67, and 92, respectively. Note that as t increases and the percentage of waiting costs becomes more significant with the respect to the other cost components, the method seems to require more CPU time. The increase in computational effort is expected as larger values of t inflate the approximation error and require the addition of more cuts.

The number of cuts generated, however, is not affected. As only a small fraction of the large number of constraints in [MIP(H)] is generated, the proposed method seems to select a small number of constraints that suffices in finding the optimal solution.

4. Conclusion

In this paper, we present a finite and exact solution method for an important problem: the service system design problem also known as the facility location problem with immobile servers, stochastic demand, and congestion. The nonlinearity of the objective is often an obstacle for the exact solution of the models provided in the literature. We exploit the structure of the nonlinear term in the objective function to provide a linearization that has a large number of constraints, and propose an exact solution method based on cutting planes. The numerical results reveal the stability and efficiency of the approach in exactly solving models with up to 100 demand points, 20 potential facilities, and 3 capacity levels.

Acknowledgments

The author thanks two anonymous referees and a senior editor for their valuable comments. The author acknowledges

Table 1 **Tests for** $\beta = 0.1; 1; 10; 100; 1{,}000$

			$\beta=0.1$					$\beta=1$					$\beta=10$					$\beta=100$					$\beta=1{,}000$				
Problem	m	n	AC[a]	FC[a]	WC[a]	CPU	Cuts	AC[a]	FC[a]	WC[a]	CPU	Cuts	AC[a]	FC[a]	WC[a]	CPU	Cuts	AC[a]	FC[a]	WC[a]	CPU	Cuts	AC[a]	FC[a]	WC[a]	CPU	Cuts
p1	50	10	32	64	4	55	4	27	59	14	47	2	16	45	39	109	3	5	23	72	279	8	0.5	4	95	223	8
p2	50	10	29	66	5	44	3	25	60	16	102	4	14	43	43	131	3	5	20	75	303	7	0.5	3	96	307	9
p3	50	10	25	71	4	69	4	21	65	14	62	2	13	50	38	98	4	5	24	71	218	5	0.5	4	95	187	6
p4	50	10	29	67	4	38	3	29	62	9	59	3	12	54	35	98	4	5	27	68	169	5	0.6	5	94	237	5
p5	50	10	30	62	8	42	4	23	52	25	82	3	9	45	46	251	3	1	24	75	361	3	0.2	3	97	735	9
p6	50	10	27	64	8	94	2	18	66	16	373	4	8	44	48	722	6	2	21	78	721	6	0.2	3	97	621	7
p7	50	10	23	70	7	106	5	18	60	23	112	3	7	50	43	466	5	1	25	73	1,072	9	0.3	3	96	470	6
p8	50	10	20	74	6	70	3	16	63	20	68	3	7	55	39	404	5	1	26	73	843	7	0.2	4	96	437	6
p9	50	10	33	63	4	49	3	27	58	15	136	5	15	41	44	203	3	3	23	74	722	6	0.3	6	94	602	5
p10	50	10	23	68	8	44	2	25	60	16	234	3	13	41	46	361	3	2	25	72	722	6	0.4	5	94	843	7
p11	50	10	25	72	3	291	5	26	56	18	93	2	12	48	41	327	4	3	26	71	1,203	10	0.4	7	93	602	5
p12	50	10	28	70	2	96	3	24	60	16	73	3	11	52	37	204	3	2	30	68	1,083	9	0.4	8	92	481	4
p13	50	20	18	73	9	361	3	17	65	19	241	2	8	53	39	602	5	2	27	70	843	7	0.3	6	94	602	5
p14	50	20	17	73	10	250	4	15	62	23	602	5	10	46	43	602	5	2	25	73	361	3	0.2	5	95	722	6
p15	50	20	20	75	5	590	6	16	67	17	482	4	9	55	36	481	4	2	26	72	602	5	0.2	6	94	843	7
p16	50	20	17	78	4	443	4	14	71	15	474	4	8	59	32	361	3	2	29	70	482	4	0.4	7	93	602	5
p17	50	20	19	69	11	424	6	18	57	25	361	3	9	52	39	481	4	2	29	70	602	5	0.2	7	93	602	5
p18	50	20	13	77	9	361	3	14	63	23	361	3	8	51	41	482	4	2	26	72	722	6	0.3	5	95	722	6
p19	50	20	20	75	5	602	5	17	65	17	602	5	7	47	46	482	4	2	29	70	1,324	11	0.2	7	93	482	4
p20	50	20	17	79	4	241	2	14	70	16	241	2	7	50	43	602	5	2	25	74	481	4	0.2	8	91	361	3
p21	50	20	23	68	9	197	4	20	63	17	361	3	12	52	36	481	4	2	27	71	481	4	0.4	8	91	602	5
p22	50	20	19	71	10	231	4	15	58	28	361	3	10	47	43	481	4	2	25	73	842	7	0.3	7	93	722	6
p23	50	20	13	80	7	481	4	18	62	20	361	3	8	59	33	1,083	9	2	26	72	1,083	9	0.2	13	87	843	7
p24	50	20	22	71	7	538	5	17	66	17	361	3	10	48	42	842	7	2	29	69	962	8	0.2	15	85	602	5
p41	90	10	34	62	4	220	5	26	68	6	97	2	18	59	23	126	2	8	30	63	167	5	1.2	10	89	167	5
p42	80	20	29	69	3	319	5	27	60	13	529	5	18	54	28	482	4	8	34	58	471	4	1.2	13	86	361	3
p43	70	30	23	72	4	603	5	32	61	8	362	3	25	42	33	603	5	7	31	62	241	2	1.3	10	88	603	5
p44	90	10	30	66	4	28	2	25	66	10	204	5	22	57	21	179	4	7	32	61	183	6	0.9	9	91	253	8
p45	80	20	24	71	4	409	4	24	63	13	602	5	19	59	22	241	2	7	39	54	603	5	1.0	13	86	362	3
p46	70	30	34	62	5	362	3	27	65	8	482	4	20	55	26	362	3	9	28	62	723	6	1.2	13	86	362	3
p47	90	10	27	66	7	62	3	30	59	11	66	2	20	56	24	109	2	6	33	61	114	4	0.9	9	90	179	7
p48	80	20	22	74	4	312	4	26	65	9	268	3	16	58	26	241	2	8	38	54	602	5	1.2	11	88	482	4
p49	70	30	25	69	6	253	3	22	65	13	241	2	16	67	17	241	2	7	26	67	482	4	1.0	16	83	482	4
p50	100	10	35	61	4	136	2	33	56	11	523	7	25	52	23	327	6	8	31	61	277	7	1.2	8	91	100	3
p51	100	20	25	68	7	146	2	34	55	11	241	2	22	52	26	241	2	10	34	57	362	3	1.3	10	89	482	4
p52	100	10	23	68	9	53	2	24	66	10	209	4	27	44	29	201	4	8	29	62	217	4	1.0	6	93	263	8
p53	100	20	25	68	7	399	5	26	58	16	482	4	22	43	36	844	7	8	33	58	603	5	1.2	8	91	362	3
p54	100	10	21	72	6	132	3	21	65	14	363	5	17	57	26	252	4	6	28	65	219	6	0.8	6	93	124	4
p55	100	20	22	70	8	395	4	28	60	12	376	4	18	46	35	815	7	6	32	61	482	4	1.0	11	88	362	3
Minimum			13	61	2	28	2	14	52	6	47	2	7	41	17	98	2	1	20	54	114	2	0.2	3	83	100	3
Average			24	70	6	245	4	22	62	15	290	3	14	51	35	400	4	4	28	67	570	6	0.6	8	92	472	5
Maximum			35	80	11	603	6	34	71	28	602	7	27	67	48	1,083	9	10	39	78	1,324	11	1.3	16	97	843	9

[a] Percentage of total cost.

partial support from the National Science and Engineering Research Council of Canada, Grant RGPIN249491-02.

Appendix

The following proposition is needed to prove the finiteness of the method.

Proposition. *At any given iteration, at least one of the n cuts associated with $R_j^{h_{\text{new}}}$, $j = 1, \ldots, n$ is new, i.e., was not generated in a previous iteration.*

Proof. Consider an iteration; say, q, where $(\bar{x}^q, \bar{y}^q, \bar{z}^q, \bar{R}^q)$ is the solution to [MIP(H^q)], the lower bound LB is different from the upper bound UB, i.e.,

$$LB < UB, \tag{A1}$$

and a new set of points

$$R_j^{h_{\text{new}}} = \frac{\sum_{i=1}^n \lambda_i \bar{x}_{ij}^q}{\sum_{k=1}^K \mu_{jk} \bar{y}_{jk}^q - \sum_{i=1}^n \lambda_i \bar{x}_{ij}^q}, \quad j = 1, \ldots, n$$

are generated. Suppose that the whole set of points $R_j^{h_{\text{new}}}$, $j = 1, \ldots, n$ was previously generated. Then,

$$R_j^{h_{\text{new}}} = \frac{\sum_{i=1}^n \lambda_i \bar{x}_{ij}^q}{\sum_{k=1}^K \mu_{jk} \bar{y}_{jk}^q - \sum_{i=1}^n \lambda_i \bar{x}_{ij}^q} \iff \sum_{i=1}^n \lambda_i \bar{x}_{ij}^q = \frac{R_j^{h_{\text{new}}}}{1 + R_j^{h_{\text{new}}}} \sum_{k=1}^K \mu_{jk} \bar{y}_{jk}^q, \quad j = 1, \ldots, n. \tag{A2}$$

Constraints (13)–(14) and the fact that $\bar{y}^q_{jk}$ is binary imply that

$$\bar{z}^q_{jk} \leq \left(\frac{(R^h_j)^2}{(1+R^h_j)^2} + \frac{1}{(1+R^h_j)^2}\bar{R}^q_j\right)\bar{y}^q_{jk}, \quad k=1,\ldots,K;\ j=1,\ldots,n.$$

Using constraint (10) and Equation (A2), we get

$$\begin{aligned}\frac{R^{h_{\text{new}}}_j}{1+R^{h_{\text{new}}}_j}\sum_{k=1}^K \mu_{jk}\bar{y}^q_{jk} &= \sum_{i=1}^n \lambda_i \bar{x}^q_{ij} = \sum_{k=1}^K \mu_{jk}\bar{z}^q_{jk}\\ &\leq \sum_{k=1}^K\left(\frac{1}{(1+R^{h_{\text{new}}}_j)^2}\bar{R}^q_j + \frac{(R^{h_{\text{new}}}_j)^2}{(1+R^{h_{\text{new}}}_j)^2}\right)\mu_{jk}\bar{y}^q_{jk},\\ &\qquad j=1,\ldots,n.\end{aligned}$$

Therefore, for all $j=1,\ldots,n$,

$$\sum_{k=1}^K\left(\frac{R^{h_{\text{new}}}_j}{1+R^{h_{\text{new}}}_j} - \frac{\bar{R}^q_j}{(1+R^{h_{\text{new}}}_j)^2} - \frac{(R^{h_{\text{new}}}_j)^2}{(1+R^{h_{\text{new}}}_j)^2}\right)\mu_{jk}\bar{y}^q_{jk} \leq 0,$$

which leads to

$$R^{h_{\text{new}}}_j \leq \bar{R}^q_j, \quad j=1,\ldots,n.$$

So,

$$\begin{aligned} LB &= \sum_{i=1}^n\sum_{j=1}^m c_{ij}\bar{x}^q_{ij} + \sum_{j=1}^m\sum_{k=1}^K f_{jk}\bar{y}^q_{jk} + t\sum_{j=1}^m \bar{R}^q_j\\ &\geq \sum_{i=1}^n\sum_{j=1}^m c_{ij}\bar{x}^q_{ij} + \sum_{j=1}^m\sum_{k=1}^K f_{jk}\bar{y}^q_{jk} + t\sum_{j=1}^m R^{h_{\text{new}}}_j\\ &= \sum_{i=1}^n\sum_{j=1}^m c_{ij}\bar{x}^q_{ij} + \sum_{j=1}^m\sum_{k=1}^K f_{jk}\bar{y}^q_{jk} + t\sum_{j=1}^m \frac{\sum_{i=1}^n \lambda_i\bar{x}^q_{ij}}{\sum_{k=1}^K \mu_{jk}\bar{y}^q_{jk} - \sum_{i=1}^n \lambda_i \bar{x}^q_{ij}}\\ &\geq \min\Bigg(\sum_{i=1}^n\sum_{j=1}^m c_{ij}\bar{x}^q_{ij} + \sum_{j=1}^m\sum_{k=1}^K f_{jk}\bar{y}^q_{jk}\\ &\qquad + t\sum_{j=1}^m \frac{\sum_{i=1}^n \lambda_i\bar{x}^q_{ij}}{\sum_{k=1}^K \mu_{jk}\bar{y}^q_{jk} - \sum_{i=1}^n \lambda_i \bar{x}^q_{ij}}, UB\Bigg) = UB,\end{aligned}$$

which means $LB = UB$: a contradiction to (A1). Hence, at any given iteration, at least one of the points $R^{h_{\text{new}}}_j$, $j=1,\ldots,n$ and its corresponding cut was not previously generated. □

Furthermore, there is a finite number of values that

$$R^{h_{\text{new}}}_j = \frac{\sum_{i=1}^n \lambda_i \bar{x}^q_{ij}}{\sum_{k=1}^K \mu_{jk}\bar{y}^q_{jk} - \sum_{i=1}^n \lambda_i\bar{x}^q_{ij}}$$

can take, as there is a finite number of possibilities for

$$(\bar{x}_{ij}, \bar{y}_{jk}) \in \left\{x_i, y_k \in \{0,1\}\colon \sum_{i=1}^m \lambda_i x_i \leq \sum_{k=1}^K \mu_{jk} y_k,\right\}$$

for all $j=1,\ldots,m$. Therefore the cutting plane algorithm is finite.

References

Amiri, A. 1997. Solution procedures for the service system design problem. *Comput. Oper. Res.* **24** 49–60.

Amiri, A. 2001. The multi-hour service system design problem. *Eur. J. Oper. Res.* **128** 625–638.

Berman, O., D. Krass. 2002. Facility location problems with stochastic demands and congestion. Z. Drezner, H. W. Hamacher, eds. *Location Analysis: Applications and Theory*. Springer-Verlag, Berlin, Germany, 329–371.

Brandeau, M. L., S. S. Chiu, S. Kumar, T. A. Grossman, Jr. 1995. Location with market externalities. Z. Drezner, ed. *Facility Location: A Survey of Applications and Methods.* Springer-Verlag, New York, 121–150.

Castillo, I., A. Ingolfsson, T. Sim. 2002. Socially optimal location of facilities with fixed servers, stochastic demand and congestion. Management Science Working Paper 02-4, School of Business, University of Alberta, Edmonton, Alberta, Canada.

Holmberg, K., M. Ronnqvist, D. Yuan. 1999. An exact algorithm for the capacitated facility location problems with single sourcing. *Eur. J. Oper. Res.* **113** 544–559.

Marianov, V., D. Serra. 1998. Probabilistic, maximal covering location-allocation models for congested systems. *J. Regional Sci.* **38** 401–424.

Marianov, V., D. Serra. 2002. Location-allocation of multiple-server service centers with constrained queues or waiting times. *Ann. Oper. Res.* **111** 35–50.

Rajagopalan, S., H. L. C. Yu. 2001. Capacity planning with congestion effects. *Eur. J. Oper. Res.* **134** 365–377.

Talim, J., Z. Liu, P. Naim, E. G. Coffman, Jr. 2001. Optimizing the number of robots for web search engines. *Telecomm. Systems* **17** 243–264.

Wang, Q., R. Batta, C. M. Rump. 2002. Algorithms for a facility location problem with stochastic customer demand and immobile servers. *Ann. Oper. Res.* **111** 17–34.