An Augmented Lagrangean Dual Algorithm for Link Capacity Side Constrained Traffic Assignment Problems

Torbjörn Larsson and Michael Patriksson

Division of Optimization
Department of Mathematics
Linköping Institute of Technology
S-581 83 Linköping
Sweden

LiTH-MAT-R-93-22

April 1, 1995

Abstract

As a means to obtain a more accurate description of traffic flows than that provided by the basic model of traffic assignment, there have been suggestions to impose upper bounds on the link flows. This can be done either by introducing explicit link capacities or by employing travel time functions with asymptotes at the upper bounds. Although the latter alternative has the disadvantage of inherent numerical ill-conditioning, the capacitated assignment model has been studied and applied to a limited extent, the main reason being that the solutions can not be characterized by the classical Wardrop equilibrium conditions; they may, however, be characterized as Wardrop equilibria in terms of a well-defined, natural generalized travel cost.

The introduction of link capacity side constraints makes the problem computationally more demanding. The availability of efficient algorithms for the basic model of traffic assignment motivates the use of dualization approaches for handling the capacity constraints. We propose and evaluate an augmented Lagrangean dual method in which the uncapacitated traffic assignment subproblems are solved with the disaggregate simplicial decomposition algorithm. This algorithm fully exploits the subproblem's structure and has very favourable reoptimization capabilities; both these properties are necessary for achieving computational efficiency in iterative dualization schemes. The dual method exhibits a linear rate of convergence under a standard nondegeneracy assumption. The efficiency of the overall algorithm is demonstrated through experiments with capacitated versions of well-known test problems, with the conclusion that the introduction of link capacities increases the computing times with no more than a factor of four.

The introduction of capacities and the algorithm suggested can be used to derive tolls for the reduction of flows on overloaded links. The solution strategy can be applied also to other types of traffic assignment models where side constraints have been added in order to refine a descriptive or prescriptive assignment model.

Keywords: Capacitated Traffic Assignment, User Equilibrium, Generalized Wardrop Conditions, Queue Equilibrium, Augmented Lagrangean, Disaggregate Simplicial Decomposition.

1 Introduction

1.1 Background and motivation

A class of mathematical models which is frequently studied in the field of urban traffic planning is the *traffic assignment* problems. These models may, depending of the characteristics of the real-world situation to be modelled and the purpose of the model, include a variety of different aspects, but they all have in common that they aim at describing, predicting or prescribing a traffic flow pattern in a road network where there is some (fixed or elastic) travel demand and where congestion effects result in flow-dependent link travel times. The flow pattern is determined according to a prescribed performance criterion which, typically, involves a measure of the disutility, e.g., cost, of the total traffic flow in the urban area. Often, the travel cost is set equal to the travel time, and these terms are therefore mostly used interchangeably.

The two most commonly employed performance criteria are the two optimality principles of Wardrop (1952). The first one is based on the intuitive behaviour of traffic, i.e., that each user of the congested traffic network seeks to minimize his/her own travel time, and it is therefore also known as the *user optimum*, or *equilibrium*, principle. If the vector of functions describing the travel times on the links of the road network is integrable and monotone, then Wardrop's first principle of optimality gives rise to a convex mathematical program (e.g., Dafermos, 1972), which was first formulated by Beckmann *et al.* (1956) for the case of separable travel cost functions. Non-integrable travel cost functions have asymmetric Jacobian matrices and the resulting traffic assignment problems are therefore called *asymmetric*. Such problems typically arise when modelling flows of different modes (e.g., Dafermos, 1972) or when link travel times depend on traffic flows on other links, and it may be formulated as, for example, a variational inequality problem (Smith, 1979, and Dafermos, 1980), a nonlinear complementarity problem (Aashtiani, 1979, and Aashtiani and Magnanti, 1981), or as a, generally nonconvex, mathematical program by the use of so called *gap functions* (e.g., Hearn *et al.,* 1984). Wardrop's second principle, the *system optimum* principle, corresponds to minimizing the total travel time. If the travel time functions are separable, monotonically increasing and convex, traffic flows in agreement with this principle can be found through the solution of a convex mathematical program.

The well-known basic model of traffic assignment has received a lot of attention and several highly efficient solution methods have been developed for it. (See Patriksson, 1994, for a thorough review of traffic assignment models and methods.) One important reason for this attention is that its simplicity and nice interpretations makes it attractive for practitioners, and that its very special structure together with its large size in practical applications makes it a challenge for academic research aiming at the development of efficient special-purpose algorithms. The reader should in particular recall that all efficient algorithms for the basic model exploit its inherent Cartesian product structure (e.g., Larsson and Patriksson, 1992).

A fundamental principle underlying the basic model is the steady-state assumption of the Wardrop conditions; thus, the model's validity and applicability rest heavily on the stability (as well as, of course, the knowledge) of its components. In a practical application it is not a trivial task to well estimate the data involved. Considering the link travel cost functions, their estimation involves the choice of a functional form and the calibration of the resulting functional parameters. The classical BPR formula [see (19)], for example, is based on the estimation of a *practical capacity*, which measures the maximal flow on a link that does not cause any significant congestion effect; the proper estimation of these capacities is, of course, not obvious. Furthermore, some quantities which highly influence the travel time on a link may vary in an unpredictable manner (e.g., the travel demand, the weather conditions and the proportion of different types of vehicles in the traffic flow); consequently, under some circumstances the model may yield traffic flows which are far from correct. If the deviation is unacceptable from the practical viewpoint, then the model needs to be refined in order to capture variations in the real-world traffic system. An example of such an improvement is the introduction of time-slices to capture variations in travel demands and travel time characteristics.

Another limitation of the basic model is that its inherent simplicity may make it inapplicable to more complex real-world traffic problems (e.g., Sender and Netter, 1970). For example, it does not capture the interactions between the flows on intersecting links, or between vehicles of different types.

Such flow relationships may be captured through the introduction of non-separable, and usually asymmetric, travel time functions. The resulting class of models has been extensively studied from a theoretical and algorithmical point of view (see, e.g., Nagurney, 1993, and Patriksson, 1994, and the references cited therein). [The asymmetric models have received a lot of attention mainly due to their mathematical elegance and nice interpretations; real-world applications are scarce, however.]

While improving the basic model's ability to accurately describe and predict a real-world traffic situation, modifications of the travel time functions are, however, not natural and adequate means for incorporating traffic flow restrictions such as link capacities, joint capacities in junctions or on two-way streets, or the presence of a traffic control policy; a fundamental reason for the inadequacy is the difficulty in estimating proper travel cost functions for describing such restrictions. The natural alternative for describing and capturing these supplementary flow restrictions is to introduce *side constraints*, which may have immediate physical interpretations. (We believe that these interpretations make it easier for a traffic engineer to identify a set of side constraints than to make proper estimates of parameter values in complex travel time functions.)

Although this approach seems to be useful from a practical point of view, it has received very little attention; the main reason for this is that the solutions to the resulting models can no longer be given characterizations as Wardrop equilibria in the classical sense. Moreover, as a result of the addition of the side constraints the Cartesian product structure of the feasible set of the basic model is lost, thus obtaining a computationally more demanding model.

We study a special side constrained model: the capacitated traffic assignment problem. The steady-state solutions to this model is known to have a characterization as a Wardrop equilibrium flow in terms of the sum of travel times and queueing delays on saturated links.

This *generalized* travel cost is, in fact, the natural one to be minimized by the individual travellers in a capacitated network with queueing. (This characterization is in Larsson and Patriksson, 1995, generalized to a class of convexly side constrained traffic assignment models, thereby showing that solutions to side constrained models have characterizations as Wardrop equilibrium flows in terms of natural cost functions.)

The objective of this paper is to establish that the capacitated traffic assignment problem may be efficiently solved through a combination of an augmented Lagrangean dualization scheme (see Bertsekas, 1982, for a comprehensive introduction to this class of methods) and an efficient solution method for the basic model, the disaggregate simplicial decomposition method (Larsson and Patriksson, 1992), thereby showing that the traffic assignment model may be computationally tractable even though it is extended with side constraints.

1.2 A review of capacitated assignment

To model congestion effects on road links many classes of travel cost functions have been suggested (see, e.g., Branston, 1976), but in practice the ones most frequently used are polynomial functions. These yield travel times that are finite for all link flows, so that the roads are implicitly modelled to be able to carry arbitrarily large volumes of traffic; in practice, however, road links of course have some finite limits on traffic flows. To cite Hearn (1980), this deficiency of the model causes that "the predicted flow on some links will be far lower or far greater than the traffic engineer knows they should be *if all assumptions of the model are correct*. In practice, the result is that the model predictions are ignored, or, more often, the user will perturb the components of the model (trip table, volume delay formulas, etc.) in an attempt to bring the model output more in line with the anticipated results." A simple way of enhancing the quality of an assignment model would thus be to include upper bounds on link flows. This can be done either explicitly, through the introduction of *link capacities*, or implicitly, through the use of *asymptotic travel time functions*, i.e., functions describing that a link's travel time goes to infinity when its flow approaches its upper bound (Daganzo, 1977a and 1977b), but neither of these two methodologies have been studied to any greater extent. It is then interesting to note that in some of the first mathematical models of traffic assignment problems (e.g., Jorgensen, 1963 and Tomlin, 1966), link flow capacity constraints were used to model congestion effects.

Link flow capacity constraints typically arise from traffic control policies or as a result of congestion. Examples of the first are speed limit regulations and cycle times of traffic signals (e.g., Yang and Yagar, 1994). These are *prescribed* capacities which are imposed upon the users of the traffic system, and they are therefore usually known exactly; they are also *hard*, in the sense that they can never be violated (unless, perhaps, by traffic offenders). [This type of capacity restrictions may cause steady-state queues to appear on the capacitated links; see Section 2.1 for a discussion on steady-state link queueing delays in capacitated networks.]

The second type of capacity restrictions is of a *descriptive* nature, and results from and varies with the prevailing traffic conditions. Under steady-state conditions, the link flows are usually much lower than the practical capacities. During peak-hours, however, the link flows are unstable and the resulting capacities (which may be taken as the estimated maximal average values of the fluctuating link flows) may be violated; when a link flow exceeds the capacity a queue is building up at the link's exit, while, during periods when the flow falls below the capacity the queue is dissolving. Obviously, no link capacity of this type is valid for every possible traffic condition; hence, the traffic model must be supplied with different capacity levels for different traffic situations (for example time-slices). [Compare with the use of different travel time functions in different time-slices, as mentioned in Section 1.1.]

Note that while the prescriptive (i.e., hard) capacity constraints of course need to be fulfilled exactly in a calculated solution, the flows resulting from a model with descriptive (i.e., weak) link capacities (which, in general, are known only approximately) may be allowed to slightly exceed their capacity restrictions. (In the latter case, the heuristic procedure outlined in Section 4.4 need not be included in the algorithm.)

From a modelling point of view, capacity constraints have the advantage of allowing the link flows to reach their upper bounds, whereas asymptotic travel time functions yield flows that are strictly below their bounds. Moreover, Boyce *et al.* (1981) have empirically found that asymptotic travel time functions yield unrealistically high travel times and devious rerouting of trips. A disadvantage of imposing explicit link capacities is that the Cartesian product structure of the uncapacitated problem is lost, thus making the problem more demanding computationally. Especially, the linear subproblem of the Frank–Wolfe and simplicial decomposition type methods will, instead of a set of shortest-route problems, become a linear multicommodity minimum cost network flow problem (Klessig, 1974), which is computationally burdensome. Under strong assumptions on the travel time functions and the choice of initial point, the multicommodity flow subproblem of the Frank–Wolfe method may be relaxed into shortest-route subproblems while maintaining convergence to an optimal flow pattern (Daganzo, 1977a and 1977b, and Hearn and Ribera, 1981). A solution to an explicitly capacitated user equilibrium assignment problem will not comply with Wardrop's first principle (Hearn, 1980); however, it will satisfy a modification of this principle where the usual travel costs are replaced by well-defined *generalized* travel costs. Computationally, the asymptotic travel time functions have the disadvantage that they may result in numerical difficulties. Also, whenever the problem is solved by a feasible-direction algorithm (e.g., the Frank–Wolfe method), these travel time functions make it necessary to initialize the algorithm through the calculation of a flow pattern which is strictly feasible with respect to the implicit upper bounds on the link flows (Daganzo, 1977b); this task is non-trivial though (e.g., Inouye, 1986).

1.3 Preview

Since an uncapacitated assignment problem may be solved very efficiently, a natural solution strategy for the capacitated model is to transform it into a sequence of uncapacitated problems, tending to one which is equivalent to the original, capacitated, problem. Most methods suggested for the capacitated model are therefore based on penalization/dualization concepts, see Section 2.2.

We employ an augmented Lagrangean dual scheme. Such schemes combine traditional exterior penalty methods with Lagrangean dual schemes; typically, they yield faster multiplier convergence than in ordinary dual schemes and also avoid the numerical ill-conditioning inherent in penalty methods. A major difference between the proposed augmented Lagrangean scheme and the ones previously studied for the capacitated model is that the uncapacitated subproblems are solved much more efficiently, using the disaggregate simplicial decomposition algorithm. This algorithm fully exploits the underlying problem structure and has very good reoptimization properties; both these facilities are of outmost importance in order to reach computational efficiency.

Because of the dual character of augmented Lagrangean schemes, feasible solutions to the original problem will, generally, be found in the limit only, even though the primal solutions' infeasibilities will in later iterations be small. We show that this weakness of augmented Lagrangean schemes can be suitably dealt with, at least in this application, through the inclusion of a heuristic procedure which constructs feasible solutions by carefully manipulating the (slightly) primal infeasible solutions to the Lagrangean subproblem.

The remainder of the report is organized as follows. In Section 2, we introduce the mathematical model of capacitated user equilibrium assignment, state the optimality conditions as a Wardrop-type principle in terms of generalized travel costs, give a queue equilibrium characterization of its solution, and review the previously suggested solution methods. We next establish a conceptual augmented Lagrangean scheme, provide convergence results, and interpret the scheme as a mathematical simulation of a real-life traffic engineering process. The fourth section describes an implementable version of the conceptual scheme. Section 5 gives computational results for small and medium-size test problems derived from well-known uncapacitated problems through the introduction of properly chosen explicit link capacities; the purpose of the experiments is to illustrate various characteristics of the scheme and prove its viability. Finally, in Section 6 we draw conclusions and suggest directions for future developments.

2 Capacitated traffic assignment

Let $\mathcal{G}$=($\mathcal{N}$,$\mathcal{A}$) denote a strongly connected transportation network, with $\mathcal{N}$ and $\mathcal{A}$ being the sets of nodes and links (arcs), respectively. For certain ordered pairs of nodes, $(p, q) \in \mathcal{C}$, where node p is the origin, node q the destination, and $\mathcal{C}$ is a subset of $\mathcal{N} \times \mathcal{N}$, there are given positive demands d_{pq} for origin-destination (or commodity) flows which give rise to a link flow pattern $f = (f_a)_{a \in \mathcal{A}}$ when distributed through the network. Associated with each link $a \in \mathcal{A}$ is a *link performance function*, $t_a : \Re_+ \mapsto \Re_{++}$, which measures the disutility of using the link as a function of its flow. Further, because of congestion effects these functions are strictly monotonically increasing. We consider the problem of determining a user equilibrium traffic flow pattern fulfilling the travel demands.

The first to clearly formulate the user equilibrium route choice criterion was Wardrop in 1952, although it already in 1920 was touched upon by Pigou.

Wardrop's first principle: The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route.

Let c_{pqr} denote the travel time on route r in origin-destination pair $(p,q) \in \mathcal{C}$ resulting from a given feasible flow pattern, and assume, without any loss of generality, that the first l routes are actually used, i.e., carry positive flows. Then, the flow pattern is a user equilibrium if and only if

$$c_{pq1} = c_{pq2} = \ldots = c_{pql},$$

and the unused routes in the origin-destination pair have travel times that are at least as large as that of the used routes, for any pair $(p,q) \in \mathcal{C}$.

Letting $\mathcal{R}_{pq}$ be the set of simple routes in origin-destination pair $(p,q) \in \mathcal{C}$, h_{pqr} the flow on route r, and π_{pq} the least travel time from node p to node q, the Wardrop user equilibrium conditions may equivalently be stated as

$$\begin{array}{rcrcrcl} h_{pqr} & > & 0 & \Longrightarrow & c_{pqr} & = & \pi_{pq}, \\ h_{pqr} & = & 0 & \Longrightarrow & c_{pqr} & \geq & \pi_{pq}, \end{array}$$

to hold for all pairs $(p,q) \in \mathcal{C}$. As was established by Beckmann *et al.* (1956), one may also derive a linearly constrained convex mathematical program whose Karush–Kuhn–Tucker conditions are equivalent to the Wardrop conditions.

2.1 The capacitated model and its optimality conditions

Introducing link capacities $u_a \in \Re_{++} \cup \{+\infty\}$, $a \in \mathcal{A}$, the arc-route formulation (e.g., Dafermos and Sparrow, 1969) of the *capacitated user equilibrium* traffic assignment problem is

[TAP-C]

$$\min \; T(f) = \sum_{a \in \mathcal{A}} \int_0^{f_a} t_a(s)ds, \tag{1a}$$

subject to

$$\sum_{r \in \mathcal{R}_{pq}} h_{pqr} = d_{pq}, \qquad \forall (p,q) \in \mathcal{C}, \tag{1b}$$

$$h_{pqr} \geq 0, \qquad \forall r \in \mathcal{R}_{pq}, \; \forall (p,q) \in \mathcal{C}, \tag{1c}$$

$$\sum_{(p,q) \in \mathcal{C}} \sum_{r \in \mathcal{R}_{pq}} \delta_{pqra} h_{pqr} = f_a, \qquad \forall a \in \mathcal{A}, \tag{1d}$$

$$f_a \leq u_a, \qquad \forall a \in \mathcal{A}, \tag{1e}$$

with

$$\delta_{pqra} = \begin{cases} 1, & \text{if route } r \text{ from node } p \text{ to node } q \text{ contains arc } a, \\ 0, & \text{otherwise} \end{cases}$$

defining the arc-route incidence matrix. Dropping the capacity side constraints (1e), the basic model of traffic assignment, to be referred to as [TAP], is obtained. We will throughout the paper presume that the link capacities are large enough to allow all travel demands to be distributed through the network, so that [TAP-C] has a feasible solution.

The problem [TAP-C] is a convex mathematical program which, because of the strict convexity of the objective with respect to the link flows, has a unique optimal link flow solution, denoted f^*, although there are, in general, alternative optimal route-flow solutions. Let π_{pq}, $(p,q) \in \mathcal{C}$, and β_a, $a \in \mathcal{A}$, denote optimal values of the Lagrange multipliers for the demand feasibility constraints (1b) and the capacity constraints (1e), respectively. Since the Abadie constraint qualification is always fulfilled for linearly constrained programs (e.g., Bazaraa *et al.*, 1993, Lemma 5.1.4), the Karush–Tuhn–Tucker conditions stated below are both necessary and sufficient for the optimality of h in [TAP-C]; similar optimality conditions have been stated by Jorgensen (1963), Hearn (1980), and Inouye (1986).

$$
\begin{aligned}
c_{pqr} + \sum_{a\in\mathcal{A}} \delta_{pqra}\beta_a &\geq \pi_{pq}, && \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \\
\beta_a &\geq 0, && \forall a \in \mathcal{A}, \\
\sum_{r\in\mathcal{R}_{pq}} h_{pqr} &= d_{pq}, && \forall (p,q) \in \mathcal{C}, \\
h_{pqr} &\geq 0, && \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \\
\sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr} &\leq u_a, && \forall a \in \mathcal{A}, \\
h_{pqr}(c_{pqr} + \sum_{a\in\mathcal{A}} \delta_{pqra}\beta_a - \pi_{pq}) &= 0, && \forall r \in \mathcal{R}_{pq},\ \forall (p,q) \in \mathcal{C}, \\
\beta_a(u_a - \sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr}) &= 0, && \forall a \in \mathcal{A}
\end{aligned}
$$

Introducing a *generalized route travel cost*

$$
\bar{c}_{pqr} = c_{pqr} + \sum_{a\in\mathcal{A}} \delta_{pqra}\beta_a, \qquad r \in \mathcal{R}_{pq},\ (p,q) \in \mathcal{C},
$$

a capacitated user equilibrium flow satisfies, for all $(p,q) \in \mathcal{C}$, the conditions

$$
\begin{array}{lclcl}
h_{pqr} > 0 & \Longrightarrow & \bar{c}_{pqr} & = & \pi_{pq}, \\
h_{pqr} = 0 & \Longrightarrow & \bar{c}_{pqr} & \geq & \pi_{pq},
\end{array}
$$

i.e., all routes utilized in an origin-destination pair have equal generalized travel costs, which are given by the optimal multiplier value π_{pq}, and no non-utilized route in the pair is cheaper. Hence, by replacing the actual travel costs with generalized ones, a capacitated user equilibrium state satisfies a Wardrop-type condition. (The converse conclusion is invalid since the complementarity conditions for the link flow capacity constraints are not necessarily satisfied whenever the above Wardrop-type conditions hold; however, a partial converse will be established below.)

The reader should note that although the equilibrium generalized route travel costs π_{pq}, $(p,q) \in \mathcal{C}$, are unique, this is not true in general for the optimal multipliers β_a, $a \in \mathcal{A}$; in the sequel, β_a, $a \in \mathcal{A}$, will denote some arbitrary optimal multiplier values.

One can, in general, not relate the actual travel costs of the unused routes to those of the used ones; for example, the cheapest route in an origin-destination pair may be unused because its generalized cost is too high. Furthermore, the Wardrop principles for [TAP] are intimately associated with the Cartesian product structure of its feasible set, and one can for the capacitated problem not state similar optimality conditions in terms of *actual* travel costs. (The extensions of Wardrop's first principle given by Anantharamaiah, 1974, and Stefek, 1989, are incorrect or, possibly, poorly formulated.) However, the following condition bears a strong resemblance to Wardrop's first principle (cf. Jorgensen, 1963, for the case of constant travel times).

Theorem 2.1 (A Wardrop-type principle) *Consider an arbitrary optimal route flow solution to* [TAP-C] *and let* $(p,q) \in \mathcal{C}$. *Assume, without any loss of generality, that the first* l *routes are actually used, and that among these the first* m *are* unsaturated, *i.e., contain no link which carries flow at its capacity level. Then,*

$$c_{pq1} = \ldots = c_{pqm} \geq c_{pq,m+1} \geq \ldots \geq c_{pql},$$

and no unused route has a lower generalized cost than the used ones.

It should be noted that the capacitated equilibrium link flow pattern found by solving [TAP-C] may also be found by solving an uncapacitated equilibrium problem [TAP] with travel time functions

$$t_a(\cdot) + \beta_a, \qquad a \in \mathcal{A}. \tag{2}$$

In the words of Jorgensen (1963), the optimal multipliers β_a, $a \in \mathcal{A}$, "measure the time gained by users of routes filled to capacity compared to the fastest route still available." Equivalently, they are the link tolls that drivers on saturated routes are willing to pay for letting them continue to use routes that are faster than the non-saturated ones. Beckmann and Golob (1974) make a similar observation for a capacitated, *system* optimum assignment model, where the multipliers are interpreted as link tolls which produce a system optimum when the individual travellers minimize their respective generalized travel costs.

In a network where saturated links have queues at their exits, the multipliers may be interpreted as the equilibrium time delays caused by the queuing. Specifically, we then assume that each link is in two distinct regimes. In the first part of the link (which includes its entrance), one observes a moving traffic stream; in the second part of the link (which includes its exit), one observes a queue whenever the link is saturated. (The queue is assumed to be short compared to the link's total length, so that the travel time in the moving stream can be considered to be unaffected by the presence of the queue.) It is then natural to consider the value $t_a(f_a)$ as being the travel time of the moving traffic stream, while the value of the multiplier term in (2) is the waiting time in the queue at the link's exit, i.e., its *queueing delay.*

In the special case when the link travel times are flow independent, Payne and Thompson (1975) [see also Smith, 1987] use the concept of queue equilibrium to establish a characterization of solutions to the capacitated problem; this result is extended to the case of non-constant link travel times by Miller *et al.* (1975).

Specifically, a feasible link flow f to the capacitated traffic assignment problem together with a vector $q \in \Re_{+}^{|\mathcal{A}|}$ of link queueing delays is defined to be a *queue equilibrium* if the links unsaturated at f carry no queues. Notice that this definition is merely a restatement of the complementarity conditions for the link flow capacity constraints. Therefore, the below characterization result is easily established.

Theorem 2.2 (A queue equilibrium characterization of solutions to [TAP-C]) *Let f be a feasible link flow solution to* [TAP-C]*. It is then an optimal link flow if and only if there is a vector β of non-negative Lagrange multipliers for the capacity constraints such that f is a Wardrop equilibrium with respect to the generalized link travel costs $t_a(\cdot) + \beta_a$, $a \in \mathcal{A}$, and (f, β) is a queue equilibrium.*

Hence, the optimal values of the Lagrange multipliers β may be interpreted as equilibrium link queueing delays.

The solution of a capacitated problem can be used as a means for guiding the traffic engineer how to correct the travel time functions in order to bring the flow pattern into agreement with the anticipated results (Hearn, 1980). As compared to heuristic adjustments of the travel time functions, the described strategy has the advantage that it is certainly easier for the engineer to give reasonable estimates of link capacities than to estimate how an adjustment of the travel time functions will affect the uncapacitated equilibrium flow pattern. The solution method to be presented can actually be interpreted as an automatized process of adjusting the travel time functions towards the correct ones, which are reached in the limit.

Moreover, in a traffic situation in which some link flows exceed desired maximal traffic volumes (which, for example, may emanate from maximal allowed concentrations of exhaust fumes or levels of noise in sensitive areas), the solution of [TAP-C] may be used for constructing price-directive traffic control schemes through link tolls given by optimal values of the multipliers β; by introducing these tolls, the flows on the overloaded links are reduced to within the desired ranges.

2.2 Review of previous solution methods

When solving [TAP] with a Frank–Wolfe type algorithm, the subproblem separates into a number of shortest route problems. However, if applying such an algorithm to [TAP-C], the subproblem becomes a linear multicommodity flow problem (Klessig, 1974), which is prohibitively expensive to solve repeatedly. Solution methods proposed for [TAP-C] are often based on the recognition of this fact; they may be divided into two categories.

In the first, attempts are made to use shortest route subproblems to generate search directions. The algorithm is initialized at an inner point with respect to the link capacities, and to ensure convergence the travel time functions must satisfy the *coercivity* condition

$$\lim_{f_a \to u_a} \int_0^{f_a} t_a(s)ds = +\infty, \qquad \forall a \in \mathcal{A}, \tag{3}$$

which effectively reduces the problem to an uncapacitated problem with asymptotic cost functions (Daganzo, 1977a and 1977b). Hearn and Ribera (1981) instead assume that the sequence $\{l_k\}$ of step lengths is bounded from below by some positive number. One sufficient condition for this assumption to be fulfilled is that the initial point is strictly better (in terms of the objective value) than any feasible solution at which some capacity constraint is active; the existence of such an initial point is not guaranteed for the travel time formulae most often used [such as the BPR formula (19)]. However, it is implied by the condition (3). A possibility to ensure convergence when using general travel time formulae is to invoke a Frank–Wolfe subproblem (a linear multicommodity flow problem) whenever the shortest route solution does not yield a sufficient progress (i.e., when a step length l_k falls below some prespecified parameter $\underline{l} > 0$); see the dissertation by Stefek (1989).

Stefek's main theme is the development of simplicial decomposition type algorithms (Hearn *et al.*, 1987) for the capacitated problem. In these algorithms, the line search step of his Frank–Wolfe type method is replaced by a multi-dimensional search over the intersection of the convex hull of the hitherto generated subproblem solutions and the set defined by the capacity constraints. The safe-guarding strategy of the Frank–Wolfe type algorithm is used also in these methods; whenever the extreme points corresponding to the shortest route patterns do not provide sufficient descent in the restricted master problem, a linear multicommodity flow subproblem is invoked. (In a direct application of simplicial decomposition, subproblems would always be multicommodity flow problems and the master problem would not include the capacity constraints; such a scheme would, however, not be efficient, because of the high computational cost of the subproblems.) Stefek also presents a variation in which Lagrange multipliers for the capacity constraints of the master problem are used to price-out those constraints in the subproblem, thereby reducing the number of iterations in which a multicommodity flow subproblem have to be invoked. Computational experiments with three medium-scale problems show that these extensions of the simplicial decomposition principle are for lightly capacitated problems superior to a straightforward application of this principle (where the multicommodity flow subproblems are solved by a Dantzig–Wolfe decomposition, i.e., a column generation, approach), but inferior for heavily capacitated ones.

In the second approach, the capacitated problem is converted into a sequence of uncapacitated problems through a penalization/dualization of the capacity constraints (1e), so that efficient methods for [TAP] may be applied for the solution of [TAP-C]. (Of course, [TAP-C] may be relaxed in alternative ways; in Hearn and Lawphongpanich, 1989, and Larsson *et al.,* 1992, the definitional constraints (1d) are Lagrangean dualized, and in Inouye, 1986, all constraints but (1d) are augmented Lagrangean dualized.)

For the case of constant travel times, Jorgensen (1963) suggests applying the Dantzig–Wolfe decomposition method (which may be interpreted as a cutting plane method applied to a dual problem), but does not give any computational results. For the case of flow-dependent travel times, he suggests using approximating piece-wise constant travel time functions; the approximate problem may then be restated as a problem with constant travel times in an enlarged network. Miller *et al.* (1975) also present a column generation approach for the case of constant travel costs, in which the restricted master problems are solved using a generalized upper bounding technique (see, e.g., Lasdon, 1970).

Hearn (1980) proposes to include the explicit link flow capacities in an extended objective function by means of an exterior penalty function, thereby obtaining an uncapacitated traffic assignment subproblem, which is solved by the Frank–Wolfe method. The behaviour of the overall penalty method is illustrated through small-size numerical examples. Inouye (1986) applies an interior penalty method in which the subproblems are solved using the Frank–Wolfe method, and presents results for a small example.

Vanderstraeten-Tilquin (1977), Hearn and Ribera (1980), and Polak (1983) all employ iterative augmented Lagrangean schemes. In Vanderstraeten-Tilquin's scheme, the uncapacitated subproblems are solved by the application of a non-linear version of the out-of-kilter method to single-commodity problems obtained in a cyclic decomposition scheme. In the algorithm of Hearn and Ribera, the subproblems are solved by the Frank–Wolfe method. They consider two types of augmented Lagrangean functions and apply one of them to a small numerical example. Vanderstraeten-Tilquin gives also two other solution principles for the capacitated problem. The first is a subgradient optimization procedure for finding optimal allocations of the total link capacities to the separate commodities; this is essentially the same algorithm as the one for linear multicommodity network flows proposed by Kennington and Shalaby (1977). The second involves the solution of a sequence of lower-dimensional subproblems obtained through partitionings of variables and relaxations of non-negativity constraints (see also, e.g., Lasdon, 1970, Chapter 5). From some experimentation with small-scale test problems, Vanderstraeten-Tilquin concludes that the latter method is unfeasible for larger problems, and that the augmented Lagrangean scheme is the most viable among the two others (at least in the absence of an *a priori* knowledge of a good estimate of the optimal objective value, which would improve the performance of the subgradient optimization scheme).

3 The augmented Lagrangean scheme

Consider the problem [TAP-C] stated as

[TAP-C]

$$\min \; T(f),$$

subject to

$$\begin{aligned} g_a(f_a) &\leq 0, \qquad \forall a \in \mathcal{A}, \\ f &\in F, \end{aligned}$$

where

$$g_a(f_a) = f_a - u_a, \qquad a \in \mathcal{A}$$

and

$$F = \left\{ f \in \Re^{|\mathcal{A}|} \;\middle|\; \exists h \text{ such that (1b)–(1d) holds} \right\}.$$

If applying an *exterior penalty method* (e.g., Fiacco and McCormick, 1968) to [TAP-C], the capacity constraints are included in an extended objective function by means of a penalty function $P : \Re^{|\mathcal{A}|} \mapsto \Re$ satisfying

(1) $P(f) \geq 0$ for all $f \in F$,

(2) $P(f) = 0$ if and only if $g_a(f_a) \leq 0$ for all $a \in \mathcal{A}$,

(3) P is continuous on F.

An example of such a penalty function is

$$P(f) = \sum_{a \in \mathcal{A}} p_a(f_a), \tag{4}$$

where

$$p_a(f_a) = r_a [g_a(f_a)]_+^{m_a} \stackrel{\text{def}}{=} r_a \max\{0, g_a(f_a)\}^{m_a}, \qquad r_a > 0,\ m_a \geq 2, \tag{5}$$

which was used by Hearn (1980) in a capacitated traffic assignment context.

Introducing a penalty parameter $c > 0$, the penalized objective

$$P_c(f) = T(f) + cP(f),$$

the penalty subproblem

$$P_c = \min_{f \in F} P_c(f),$$

which amounts to solving an uncapacitated traffic assignment problem, and its solution

$$f(c) = \arg \min_{f \in F} P_c(f),$$

one may show that

(1) $P_c \leq T(f^*)$ for all $c > 0$, and

(2) $\lim_{c \to +\infty} f(c) = f^*$.

For a differentiable and separable penalty function, like (4), optimal Lagrange multipliers for the penalized constraints may be estimated using the result (e.g., Hearn, 1980)

$$\lim_{c \to +\infty} c \left. \frac{\mathrm{d}p_a}{\mathrm{d}f_a} \right|_{f_a = f_a(c)} = \beta_a, \qquad \forall a \in \mathcal{A}.$$

In order to avoid the ill-conditioning inherent in the penalty approach, one may introduce a Lagrangean term in the extended objective, thus creating an *augmented Lagrangean function* (Hestenes, 1969, Powell, 1969, Rockafellar, 1973b, and Bertsekas, 1975 and 1982). Letting μ denote the vector of Lagrange multipliers for the dualized constraints and using the penalty function (4) with $r_a = \frac{1}{2}$ and $m_a = 2$, $a \in \mathcal{A}$, the augmented Lagrangean function becomes (Rockafellar, 1973a)

$$L_c(f, \mu) = T(f) + \sum_{a \in \mathcal{A}} \overline{p}_a(f_a, \mu_a, c), \tag{6}$$

with

$$\overline{p}_a(f_a, \mu_a, c) = \frac{1}{2c}([\mu_a + c g_a(f_a)]_+^2 - \mu_a^2), \qquad a \in \mathcal{A}. \tag{7}$$

Defining the augmented Lagrangean dual objective function through the solution of the uncapacitated traffic assignment subproblem

$$L_c(\mu) = \min_{f \in F} L_c(f, \mu) \tag{8}$$

and denoting the subproblem solution by

$$f(\mu, c) = \arg\min_{f \in F} L_c(f, \mu), \tag{9}$$

we have that (Rockafellar, 1973a), for any $c \geq 0$,

(1) $L_c(\mu) \leq L_c(\beta) = T(f^*)$ for all $\mu \geq 0$,

(2) $\lim_{\mu \to \beta} f(\mu, c) = f(\beta, c) = f^*$.

Hence, the augmented Lagrangean dual objective function is, for any $c \geq 0$, maximized by arbitrary optimal values of the Lagrange multipliers, and the optimal flow pattern may be obtained for finite values of the penalty parameter. Also, although the flow pattern $f(\mu, c)$ is in general infeasible in [TAP-C] unless $\mu = \beta$, it will become near-feasible for near-optimal values of the multipliers.

The choice $c = 0$, which gives the ordinary Lagrangean dualization scheme, is feasible because of the strict convexity of T; see, e.g., the discussion following Theorem 6.5.1 in Bazaraa *et al.* (1993). In general, however, the augmented Lagrangean schemes have superior convergence characteristics, and from now on, we thus presume that $c > 0$.

Optimal multipliers may be found by solving the augmented Lagrangean dual problem

$$\max_{\mu} L_c(\mu), \tag{10}$$

where L_c is concave and differentiable, with

$$\frac{\partial L_c(\mu)}{\partial \mu_a} = \max\left\{ g_a(f_a(\mu, c)), -\frac{\mu_a}{c} \right\}, \qquad a \in \mathcal{A}. \tag{11}$$

A steepest ascent multiplier update with step length c yields (see Bertsekas, 1982, p. 162)

$$\mu_a := [\mu_a + c g_a(f_a(\mu, c))]_+, \qquad a \in \mathcal{A}; \tag{12}$$

if c is sufficiently small, the value of L_c will ascend. [One may also show (Bertsekas, 1982, Proposition 5.8) that if μ is sufficiently close to an optimal dual solution, also the value of the Lagrangean dual function L_0 will always ascend.]

Although convergence is ensured for any positive value of c, a good practical performance demands for a careful choice (e.g., Hestenes, 1975, and Bertsekas, 1982). Especially, there is a trade-off between a high rate of convergence in the multiplier space and the degree of ill-conditioning of the Lagrangean subproblem; see Luenberger (1984, Chapter 13). Usually, the initial value of c chosen is small, and then increased whenever a measure of the total infeasibility in the dualized constraints does not improve sufficiently fast (e.g., Powell, 1969). We thus introduce a non-decreasing sequence $\{c_k\}$ of positive penalty parameters, and define a sequence of primal-dual iterates through the formulae

$$f^k = f(\mu^k, c_k), \tag{13}$$
$$\mu_a^{k+1} = [\mu_a^k + c_k g_a(f_a^k)]_+, \qquad a \in \mathcal{A}, \tag{14}$$

where $k = 1, 2, \ldots$, with μ^1 being some initial guess.

Theorem 3.1 (Convergence of the augmented Lagrangean scheme) *Let μ^1 be arbitrary, $\{(f^k, \mu^k)\}$ be given by (13), (14), and $\{c_k\}$ satisfy $c_k \geq c > 0$ for all k. Then, $\{f^k\} \to f^*$ and every accumulation point of $\{\mu^k\}$ (at least one exists) is a vector of optimal Lagrange multipliers for (1e).*

Proof. Follows from, e.g., Kort and Bertsekas (1976). □

Remark 3.1 If $c_k \equiv c > 0$, then the sequence $\{\mu^k\}$ converges; furthermore, this is true even when the subproblem (13) is solved inexactly only (Rockafellar, 1973b, 1976a).

Remark 3.2 As stated in Section 2.1, the optimal Lagrange multipliers for the capacity constraints may be seen as link tolls which, when imposed upon the travellers, yield an uncapacitated user equilibrium traffic flow pattern that fulfils the link capacities. The iterative search procedure (13), (14) may thus be interpreted as a mathematical *simulation* of a real-life process in which a traffic engineer attempts to limit link flows by introducing link tolls and modifying them until the travellers' behavioural response is the intended one. Moreover, the traffic engineer employs the very natural strategy of modifying the link tolls in proportion to the violations of the link flow limitations that he/she tries to impose. (This strategy for finding suitable link tolls can certainly not be implemented in the real-life traffic system.)

To be able to obtain rate of convergence results, one must impose a strict complementarity condition which amounts to the assumption that a user of a saturated link is faced with an additional cost, for example delay in a queue (see Section 2.1).

Strict complementarity assumption. Let (f^*, β) be any optimal primal-dual pair. Then $\beta_a > 0$ for any $a \in \mathcal{A}$ such that $f_a^* = u_a$.

Theorem 3.2 (Convergence rate results for the augmented Lagrangean scheme) *Let $t_a : \Re_+ \mapsto \Re_{++}$ be continuously differentiable on $\Re_+$ for all $a \in \mathcal{A}$, and let the strict complementarity assumption hold. Let μ^1 be arbitrary, $\{(f^k, \mu^k)\}$ be given by (13), (14), and $\{c_k\}$ satisfy $c_{k+1} \geq c_k \geq c > 0$ for all k.*

(a) *$\{f^k\} \to f^*$ and $\{\mu^k\}$ converges to a vector β of optimal Lagrange multipliers for (1e). Furthermore, for all $a \in \mathcal{A}$ such that $f_a^* < u_a$ holds $\{\mu_a^k\}$ converges to zero finitely.*

(b) *If $\limsup_{k\to\infty} c_k < +\infty$ and $\mu^k \neq \beta$ for all k, then $\{\mu^k\}$ converges linearly.*

(c) *If $\{c_k\} \to +\infty$ and $\mu^k \neq \beta$ for all k, then $\{\mu^k\}$ converges superlinearly.*

Proof. The theorem follows from results in Bertsekas (1982). Under the assumptions on [TAP-C], Assumption (S^+) on p. 161 holds. Then, from Proposition 3.2, this is also true for Assumption (S) on p. 104.

(a) Follows from Theorem 3.1 and Bertsekas (p. 162).

(b) Follows from Proposition 2.7 and the above.

(c) Follows from Proposition 2.7 and the above. □

Remark 3.3 Corresponding results for inexact solutions of the subproblem (13) can be derived from results in Kort and Bertsekas (1976), Bertsekas (1982) and Rockafellar (1976a).

Remark 3.4 Rockafellar (1976a and 1976b) shows that the convergence of the augmented Lagrangean algorithm is finite if the objective exhibits a sharpness property (Polyak, 1987, p. 136). In the dually equivalent proximal point algorithm, finite convergence is achieved under the slightly less restrictive weak sharpness assumption (Ferris, 1991), and, thus, the augmented Lagrangean algorithm is finite whenever (10) has a weak sharp solution. Sufficient conditions for the solution to be weak sharp is a subject for further research.

Remark 3.5 The strong relationships between the augmented Lagrangean algorithm, the proximal method of multipliers (e.g., Rockafellar, 1976a), and the class of regularization methods (e.g., Polyak, 1987, Section 6.1.2) allow us to strengthen the convergence results for the sequence $\{\mu^k\}$: under the assumptions of Theorem 3.2.c, the sequence $\{\mu^k\}$ converges to the vector of optimal multipliers of minimal Euclidean norm. The simulation of the traffic engineer's strategy outlined in Remark 3.2 thus automatically yields the link tolls which are minimal (in the sense of the Euclidean norm). A similar nice interpretation is obtained when viewing the optimal multipliers as equilibrium queueing delays.

In general, the sequence of primal iterates generated by the augmented Lagrangean scheme will converge only asymptotically to the optimal flow pattern, and feasible solutions to [TAP-C] are not found finitely. We will therefore in the final algorithm include a procedure that heuristically converts the (approximate) subproblem solutions into feasible solutions to [TAP-C]; see Section 4.4.

4 An implementable version of the scheme

We now give our implementations of the steps of the conceptual scheme described so far. Finally, the complete algorithm is summarized along with its convergence characteristics.

4.1 Initialization of the Lagrange multipliers

The choice of initial multiplier values is crucial to the overall algorithm efficiency (e.g., Bertsekas, 1982, Section 2.2.5), and it is advisable to utilize any knowledge about the problem's properties to find values that are believed to be near-optimal.

Denote by f^0 the solution to [TAP], let $\overline{\mathcal{A}}^0 = \{ a \in \mathcal{A} \mid f_a^0 > u_a \}$ and $\overline{t}_a = t_a(f_a^0) - t_a(u_a)$, $a \in \overline{\mathcal{A}}^0$ (see Figure 1).

[Place of Figure 1]

To motivate our choice of initial multiplier values, we consider a single over-saturated link $a \in \overline{\mathcal{A}}^0$. By introducing the link toll $\overline{t}_a$, the travellers will already at a flow at the capacity level face a generalized link travel cost that equals the actual travel cost that he/she was willing to pay before the link toll was introduced. A portion of the travellers which corresponds to the excess flow will therefore take other routes, so that the link flow becomes feasible with respect to the capacity restriction and also satisfies the generalized Wardrop condition. We thus choose the values

$$\mu_a^1 = \begin{cases} t_a(f_a^0) - t_a(u_a), & \text{if} \quad a \in \overline{\mathcal{A}}^0, \\ 0, & \text{if} \quad a \notin \overline{\mathcal{A}}^0, \end{cases} \tag{15}$$

which are believed to be near-optimal. However, they are not likely to be optimal since the excess link flow rerouting will alter the conditions on the other links.

4.2 Choice of penalty parameter values

Guidelines for the choice of the sequence $\{c_k\}$ of penalty parameters are given, for example, in Bertsekas (1982, Section 2.2.5), but a good choice usually requires some experiments. In our application, convergence is guaranteed for any positive value of the penalty parameter, and ill-conditioning of the subproblems (13) may therefore be avoided. On the other hand, the values must be large enough to give a high rate of convergence of the multiplier iteration.

We have used the updating rule suggested by Bertsekas (1982, p. 123), in which, at each iteration, the penalty parameter is increased if the total infeasibility in the dualized constraints does not decrease sufficiently. With $g_a^+(f_a, \mu_a, c) = \max\{g_a(f_a), -\mu_a/c\}$, we let

$$c_{k+1} = \begin{cases} \kappa c_k, & \text{if } \|g^+(f^k, \mu^k, c_k)\| > \gamma \|g^+(f^{k-1}, \mu^{k-1}, c_{k-1})\|, \\ c_k, & \text{otherwise.} \end{cases} \tag{16}$$

Bertsekas recommends choosing the parameter values $2 \le \kappa \le 10$ and $\gamma = 0.25$.

4.3 Solving the subproblems

The uncapacitated traffic assignment subproblems that need to be solved during the course of the dual scheme differ with respect to the objective functions only, and it is therefore suitable to choose a method with reoptimization features for their solution. A traffic assignment method with particularly good reoptimization facilities is the *disaggregate simplicial decomposition* (DSD) algorithm (Larsson and Patriksson, 1992), which works as follows as applied to the subproblem (13).

The key observation behind the algorithm is that the feasible set of the basic model of traffic assignment is a Cartesian product with respect to origin-destination pairs, provided that the auxiliary link-flow defining constraints (1d) are handled implicitly. If, in the application of a simplicial decomposition scheme, each of these sets is represented separately, the disaggregate version of the scheme is obtained. In its master problem, there is one convexity constraint for each origin-destination pair $(p, q) \in \mathcal{C}$, and each convexity variable defines the portion of the origin-destination demand d_{pq} which is distributed along a specific route in the pair. Assuming that nonempty subsets $\hat{\mathcal{R}}_{pq}$ of the sets $\mathcal{R}_{pq}$ of simple routes, $(p, q) \in \mathcal{C}$, are known, denoting by λ_{pqr} the variable corresponding to route r in $\hat{\mathcal{R}}_{pq}$ and by λ the vector of all variables, the total link flows are

$$f_a(\lambda) = \sum_{(p,q)\in\mathcal{C}} d_{pq} \sum_{r\in\hat{\mathcal{R}}_{pq}} \delta_{pqra}\lambda_{pqr}, \qquad a \in \mathcal{A},$$

and the restricted disaggregated master problem is given by

$$\min_{\lambda} \; L_c(f(\lambda), \mu),$$

subject to

$$\begin{aligned} \sum_{r\in\hat{\mathcal{R}}_{pq}} \lambda_{pqr} &= 1, \qquad \forall (p, q) \in \mathcal{C} \\ \lambda_{pqr} &\ge 0, \qquad \forall r \in \hat{\mathcal{R}}_{pq}, \; \forall (p, q) \in \mathcal{C}. \end{aligned}$$

The solution of the restricted master problem provides an upper bound on the value $L_c(\mu^k)$. The algorithm proceeds by linearizing the objective $L_c(\cdot, \mu^k)$ with respect to the link flow variables at the solution produced by the restricted master problem, and solving the resulting linearized version of (13), which amounts to calculating the shortest routes for all origin-destination pairs. The sets $\hat{\mathcal{R}}_{pq}$ are then augmented by the routes not already contained in the sets. Since $L_c(\cdot, \mu^k)$ is convex, the solution of the linearized problem provides a (Frank–Wolfe) lower bound on $L_c(\mu^k)$. The greatest lower bound found hitherto is denoted LBD. The procedure iterates until the relative difference between the upper and lower bounds on the value $L_c(\mu^k)$ is small enough, and, at termination, the solution to the latest restricted master problem is an approximate augmented Lagrangean subproblem solution, f^k. It is important to notice that the value LBD will at termination provide a lower bound on the optimal value of [TAP-C], since the latter is bounded from below by $L_c(\mu^k)$; see Section 3.

The master problem is a convex program with very simple linear constraints. In Larsson and Patriksson (1992), it is solved using a scaled reduced gradient method whose line search is performed in total link flows using the Armijo step length rule. The validity of the disaggregate simplicial decomposition algorithm is a consequence of the proof given by von Hohenbalken (1977), although his approach includes the dropping of all columns with zero weights. The very good reoptimization capabilities of the disaggregate simplicial decomposition algorithm is due to the storage of routes, which enables easy reoptimization with respect to changes in link performance functions, travel demands and network topology, through appropriate modifications of the latest restricted master problem and its solution.

4.4 Generating feasible solutions

A heuristic procedure for converting an approximate subproblem solution, f^k, into a feasible solution to [TAP-C], denoted $\overline{f}^k$, should fulfil two requirements. First, in order to find the optimal link flows in the limit, the heuristic alteration of the subproblem solution should be conservative in the following sense. Let $\overline{f} = Pr(f)$ be a heuristic projection of $f \in F$ onto the feasible set of [TAP-C], i.e., onto $F_C \stackrel{\text{def}}{=} \{\, f \in F \mid f \leq u \,\}$. If the mapping Pr has the property

$$\|Pr(f) - f\| \to 0 \text{ when } \min_{y \in F_C} \|y - f\| \to 0, \tag{17}$$

then

$$\begin{aligned} \|\overline{f}^k - f^*\| &\leq \|\overline{f}^k - f^k\| + \|f^k - f^*\| \\ &= \|Pr(f^k) - f^k\| + \|f^k - f^*\| \\ &\to 0, \text{ when } k \to \infty, \end{aligned}$$

i.e., the sequence $\{\overline{f}^k\}$ of feasible solutions tends to f^*. Second, in order to make the heuristic procedure computationally cheap, the structure of the feasible set must be exploited in its construction.

The idea behind our feasibility heuristic, which is similar to that in Larsson and Liu (1989) for the linear multicommodity network flow problem, is to reduce the flows on the links which are over-saturated at some main iteration k, denoted $\overline{\mathcal{A}}^k = \{a \in \mathcal{A} \mid f_a^k > u_a\}$, by repeatedly shifting flow from a route in an origin-destination pair utilizing over-saturated links to routes within the same pair that are strictly feasible with respect to the capacities; these shifts of flows will clearly maintain the feasibility with respect to the travel demands.

The origin-destination pairs are selected cyclically. Within a pair (p, q), we find routes $r, s \in \hat{\mathcal{R}}_{pq}$ with some link in $\overline{\mathcal{A}}^k$ and with all links strictly feasible with respect to the capacities, respectively. A commodity flow is then shifted from route r to route s so that either the flows on the links defining route r satisfy their respective capacities, a link contained in route r is emptied of flow, or some link along route s becomes saturated. Then, the set $\overline{\mathcal{A}}^k$ is updated. The procedure is repeated until all flows on links along the routes in $\hat{\mathcal{R}}_{pq}$ satisfy the link capacities or no flow can be shifted within the pair (p, q), and then another pair is selected. The process terminates when all link flows satisfy their respective capacities, i.e., $\overline{\mathcal{A}}^k = \emptyset$, or when no flow shift is possible between any two routes in $\hat{\mathcal{R}}_{pq}$ for any pair $(p, q) \in \mathcal{C}$. In the former case, a feasible flow, $\overline{f}^k$, has been constructed, giving an upper bound on the optimal value of [TAP-C]. In the latter case, the heuristic has failed. The lowest upper bound found so far is denoted UBD.

Clearly, the total amount of infeasibility decreases monotonically each time a route flow is shifted, and each of these shifts involve at most $|\mathcal{N}|$ link flows. Hence,

$$\sum_{a \in \mathcal{A}} \left| \overline{f}_a^k - f_a^k \right| \leq |\mathcal{N}| \sum_{a \in \overline{\mathcal{A}}^k} (f_a^k - u_a),$$

so that $\|\overline{f}^k - f^k\| \to 0$ when $\{f^k\} \to f^*$, and we conclude that the sequence $\{\overline{f}^k\}$ of feasible solutions is optimizing in the limit (provided, of course, that the heuristic is successful in an infinite number of iterations). Also, this heuristic procedure is computationally cheap and easily implemented since the route flows are explicitly available from the disaggregate simplicial decomposition scheme.

4.5 Termination criteria

When the feasibility heuristic succeeds, both lower and upper bounds on the optimal value are available, and a natural termination criterion is

$$\frac{UBD - LBD}{LBD} < \varepsilon_1,$$

where $\varepsilon_1 > 0$ is a prespecified parameter.

If the heuristic fails to generate upper bounds (which may, in particular, happen in the early iterations), we employ a safe-guard termination criterion based on a measure of the infeasibility of (f^k, μ^k) with respect to the complementary slackness conditions

$$\mu_a(f_a - u_a) = 0, \qquad \forall a \in \mathcal{A};$$

then the algorithm is terminated if

$$e_{\max} = \max_{a \in \mathcal{A}} \left\{ \frac{|f_a^k - u_a|}{u_a} \;\middle|\; \mu_a^k > 0 \right\} < \varepsilon_2, \tag{18}$$

where $\varepsilon_2 > 0$ is a prespecified parameter.

4.6 Summary of the implemented scheme

We summarize below the steps of the algorithm.

Step 0 (*Initialization*) Solve [TAP] approximately using the DSD algorithm, giving f^0 and the lower bound LBD. Choose μ^1 according to (15), and $c_1 > 0$. Choose $\varepsilon_1, \varepsilon_2 > 0$. Let $k = 1$.

Step 1 (*Subproblem solution*) Reoptimize the augmented Lagrangean subproblem (13) approximately using the DSD algorithm, starting from f^{k-1} and giving f^k. Update the lower bound LBD.

Step 2 (*Generation of feasible solution*) Generate, if possible, a feasible solution $\overline{f}^k$ from f^k using the heuristic described in Section 4.4. If the heuristic succeeds, then calculate the upper bound UBD.

Step 3 (*Convergence check*) If $\frac{UBD-LBD}{LBD} < \varepsilon_1 \rightarrow$ Stop. If an upper bound is not available, then calculate $e_{\max}$ according to (18). If $e_{\max} < \varepsilon_2 \rightarrow$ Stop. Otherwise, continue.

Step 4 (*Multiplier and penalty update*) Let μ^{k+1} be given by (14) and c_{k+1} by (16). Let $k := k + 1$, and go to Step 1.

The convergence characteristics of the algorithm may be summarized as follows.

Theorem 4.1 (Convergence of the overall scheme) *Let $\mathcal{S}$ denote the sequence of iterations in which the primal feasibility heuristic is successful, and assume that it is infinite. Under the assumptions stated previously, the algorithm generates sequences $\{\mu^k\}$, $\{c_k\}$, $\{f^k\}$, and $\{\overline{f}^k\}_{k\in\mathcal{S}}$ which are bounded and fulfil*

(1) $\{\mu^k\} \rightarrow \beta$ *(linearly), and* $\{\mu_a^k\} \rightarrow 0$ *(finitely) for all* $a \in \mathcal{A}$ *such that* $f_a^* < u_a$*,*

(2) $\{f^k\} \subset F$ *and* $\{f^k\} \rightarrow f^*$*,*

(3) $\{\overline{f}^k\}_{k\in\mathcal{S}} \subset F_C$ *and* $\{\overline{f}^k\}_{k\in\mathcal{S}} \rightarrow f^*$*.*

Proof. Follows from results of Bertsekas (1982, Section 2.2.5 and Corollary 5.9), Theorem 3.2, and the result derived in Section 4.4. □

5 Computational study

In order to investigate the efficiency of the proposed algorithm, it was coded in double precision FORTRAN-77 on a SUN 4/390 computer and numerically tested on a number of test networks, constructed from uncapacitated test networks from the literature. In Table 1 we give the origins and sizes of the networks used.

[Place of Table 1]

All test problems employ the travel time formula of the Bureau of Public Roads (1964),

$$t_a(f_a) = t_a^0 \left(1 + 0.15 \left(\frac{f_a}{c_a}\right)^{m_a}\right), \qquad a \in \mathcal{A}, \tag{19}$$

where t_a^0 is the free-flow travel time on link a, $m_a \geq 1$, and c_a is its *practical capacity.* The link capacities u_a were chosen as

$$u_a = K c_a, \qquad K > 0, \qquad \forall a \in \mathcal{A}, \tag{20}$$

i.e., as a uniform scaling of the practical capacities. For each of the problems, the constant K, which we will refer to as the *capacity scaling factor*, was chosen as small as possible without causing infeasibility; thus, difficult instances of [TAP-C] are created. (Note that, through the relations (1b), (1d) and (19), a uniform downward scaling of the vector c of practical capacities is equivalent to a uniform upward scaling of the demand vector d. Hence, our choice of capacity scaling factor corresponds to uniformly increasing the travel demands until the link capacities do not allow any more travellers through the network.)

5.1 Implementational considerations

In the penalty parameter updating formula (16), the values $\kappa = 5$ and $\gamma = 0.25$ worked well for all test problems. The initial value of the penalty parameter, c_1, was chosen such that the Lagrangean and penalty terms in (6) had the same magnitude. (A much larger initial value degrades the rate of convergence because of numerical ill-conditioning, while a very small initial value results in poor convergence of the multiplier iteration.)

In the disaggregate simplicial decomposition algorithm, the value of the acceptance parameter in the Armijo step length rule was chosen as in the experiments described in Larsson and Patriksson (1992), i.e., 0.2 for the smaller problems and 0.3 for the larger ones. Also, the solution of a restricted disaggregated master problem was terminated as in those experiments. The relative accuracy demanded for each augmented Lagrangean subproblem (13) was, for most problems, 1.0%. This accuracy was sufficient to reach a solution within 0.1% to 1.0% of the optimal value of [TAP-C]. When demanding solutions of higher accuracies, it was also necessary to solve the subproblems more accurately. The strategy chosen in that case was to demand a relative accuracy of 1.0% initially, and then divide it by two for three successive iterations.

5.2 Computational results

The proposed method may be viewed as a combination of Lagrangean duality and penalty approaches, which is intended to inherit these approaches' positive characteristics while avoiding their respective negative ones, and it is therefore of interest to compare its performance to that of pure Lagrangean and penalty schemes, respectively. To verify the superiority of the combined scheme, we made a preparatory experiment in which we compared the three methods' performance on the Sioux Falls test network. The conclusion was that the augmented Lagrangean scheme indeed outperforms the other two methods; see Larsson and Patriksson (1994) for a detailed description of these experiment.

In a second preparatory experiment, we investigated how the properties of a capacitated traffic assignment and the behaviour of the augmented Lagrangean scheme vary with the tightness of the link capacities. This was done by solving the test network of Hearn and Ribera for various values of the capacity scaling factor, i.e., by uniformly rescaling the capacities. As expected, the computing times are increasing with decreasing capacities; this was also the case for the average generalized travel times for the utilized routes. However, it was observed that the average actual travel times for the utilized routes may sometimes decrease with decreasing capacities; this is due to the phenomenon known as *Braess' paradox* (e.g., Sheffi, 1985, pp. 75–77). To establish how the computational difficulty of the augmented Lagrangean subproblems varied, the average numbers of routes generated and utilized were also recorded. These figures are of interest since they give a notion of the size and difficulty of the restricted master problems. As expected, the numbers of routes generated and utilized increase rapidly for small values of the capacity scaling factor, indicating higher congestion. Moreover, when the problem is tightly capacitated, almost all routes generated are also used. Details about this experiment can also be found in Larsson and Patriksson (1994).

In the remainder of this section, computational results for the networks in Table 1 are reported. To enable conclusions about the true deviations from optimality at termination to be drawn, we first solved the test problems demanding very high accuracies and recorded the optimal values. For each problem, we give the number of link flows that are initially over-saturated and saturated at termination, respectively, and the computing time needed to obtain a given accuracy. We also present the number of routes generated and utilized.

The first test problem is a small-size network with a quadratic objective (i.e., $m_a = 1$ for all $a \in \mathcal{A}$), for which the capacity scaling factor $K = 5.5$ was chosen. Initially, three links were over-saturated. At termination, after four iterations, a feasible solution with an objective value of 85757.52 and a relative error of 0.37% is obtained. The true deviation is, however, less than 0.0055%. The computing time was 0.87 second and the average numbers of roads generated and utilized within the four origin-destination pairs were 2.75 and 2.25, respectively. At termination, two links out of the initially three over-saturated ones were saturated.

The second problem is also a small network with a quadratic objective. With $K = 1.5$, two of the links were over-saturated initially. After five iterations and 0.60 cpu second a feasible solution with objective value 1483.22 and the relative error 0.49% was obtained. The true deviation from optimum was less than 0.041%, and the average number of routes generated and utilized were 6.75 and 5.75, respectively. Both of the initially over-saturated links were saturated at termination.

The third problem differs from the second one in travel time formulae only (in particular, $m_a = 4$ for all $a \in \mathcal{A}$). In Hearn and Ribera, capacities corresponding to $K = 1$ are used.

They apply an augmented Lagrangean scheme with

$$\overline{p}_a(f_a, \mu_a, c) = \begin{cases} \mu_a g_a(f_a) + \frac{c}{2} g_a^2(f_a), & \text{if } \mu_a > 0, \\ \frac{c}{2} g_a(f_a) \left(\frac{g_a(f_a) + |g_a(f_a)|}{2} \right), & \text{if } \mu_a = 0, \end{cases}$$

and in which the subproblems are solved by the Frank–Wolfe method. This augmented Lagrangean function, which was first given by Pierre and Lowe (1975), was tested in our implementation; the result was very similar to that of using (7). After 25 iterations and 2.9 seconds of computing time our method gave a dual objective value of 2307.91. The feasibility heuristic failed in all iterations and the termination was therefore based on the criterion (18). In Table 2, we present the link flows produced by the algorithm of Hearn and Ribera (column A), and the proposed one (column B), respectively.

[Place of Table 2]

The fourth problem models the traffic in the city of Sioux Falls, South Dakota, using link performance functions of the form (19), with $m_a = 4$ for all $a \in \mathcal{A}$. Here, $K = 2.0$ was used. After two iterations the algorithm terminated with an upper bound of 43.371. The relative error was 0.43% while the true deviation from the optimum is less that 0.22%. The computing time used was 7.6 seconds and the average number of routes generated and utilized per origin-destination pair was 2.23 and 1.47, respectively. Initially, 14 links were over-saturated, and, at termination, all of these were saturated. In Figure 2 the upper (solid line) and lower bounds (dashed line) are given.

[Place of Figure 2]

The fifth network models the city of Winnipeg and employs travel time formulae with different values of m_a. The scaling factor $K = 1.8$ was used. After 776 cpu seconds and 12 iterations an objective value of 892201.2 was found. The relative error was 0.72%. The feasibility heuristic succeeded in finding a feasible flow in the last iteration only. On average, 5.19 routes were generated in each origin-destination pair, and 3.34 of these were utilized. Initially, 12 links were over-saturated, and finally, all of these were saturated.

6 Conclusions and further research

6.1 Capacitated assignment

We have shown that the link capacity side constrained traffic assignment model is computationally tractable through the use of an augmented Lagrangean dual scheme. Some technical contributions are also made. First, under a weak regularity assumption, we establish a linear rate of convergence of the sequence of multiplier iterates produced by the algorithm. Second, an advanced choice of initial multiplier values is proposed, and, third, we give a procedure that constructs feasible flow patterns by carefully manipulating the subproblem solutions, which are, in general, infeasible in the original problem.

Our experiments demonstrate that the capacitated model requires a computing time which is, at most, a factor of four greater than that for the uncapacitated model solved in the initialization of the dual scheme; we consider this increase to be quite modest taking into account that the product structure of the basic model of traffic assignment has been lost.

The initialization of the Lagrange multipliers according to the expression (15) is justified by the observation that it leads to significant reductions in the excess flows on the initially over-saturated links. As a rule of thumb, we suggest choosing the initial value of the penalty parameter so that the penalty and Lagrangean terms of the augmented Lagrangean function have the same magnitude. However, the rate of convergence is acceptable also when the penalty term is small compared with the Lagrangean term. In general, only few iterations were needed to obtain a good accuracy; this was expected because of the generally rapid multiplier convergence in augmented Lagrangean schemes.

Whenever the heuristic procedure succeeds in finding a feasible solution, the deviation of the final upper bound from optimality is usually much smaller than the relative accuracy given by the algorithm; this is because of the poor quality of the linearization based lower bound. (In Larsson and Patriksson, 1992, it was established that the quality of the linearization bound is, for the uncapacitated traffic assignment problem [TAP], in general rather poor, and this property is inherited by the subproblem of the augmented Lagrangean method.) However, the lower bound of the augmented Lagrangean method has been observed to be much stronger than the one provided by an ordinary Lagrangean relaxation scheme (Larsson and Patriksson, 1994).

The efficiency and very good reoptimization capabilities of the disaggregate simplicial decomposition method (Larsson and Patriksson, 1992) motivated its use for solving the sequence of uncapacitated traffic assignment subproblems. Indeed, the computational effort needed for solving the subproblems was observed to decrease significantly for every iteration of the augmented Lagrangean method. Clearly, there is a trade-off between the accuracy to which the subproblems are solved and the number of iterations that are needed to reach convergence in the overall scheme; our experience is that it is not worthwhile to demand a high accuracy when solving the subproblems, at least not in the early iterations.

Although the proposed algorithm's practical performance is quite acceptable, it can be enhanced in various respects. First, the augmented Lagrangean scheme may be improved in several ways, for example through the use of more advanced multiplier iteration formulae (see, e.g., Section 2.3 of Bertsekas, 1982) or other types of augmented Lagrangean functions (e.g., Tseng and Bertsekas, 1993) than the one used in our development. The algorithm can also be improved by employing alternative updating formulae for the penalty parameter (e.g., Bertsekas, 1982), or by penalizing constraints using different, and individually updated, parameters. Another possibility is to scale the constraints, for example so that their right hand sides become equal, before dualizing them; such a prescaling corresponds to introducing individual penalty parameters which are however then uniformly updated (Conn *et al.*, 1991). (It may, alternatively, be interpreted as a preconditioning of the dual objective function through the transformation of the multiplier space with a diagonal matrix.) Second, the heuristic procedure for generating feasible solutions to [TAP-C] from the subproblem solutions can be designed differently. In particular, a more effective procedure may be devised by manipulating the link flows instead of the route flow solution, and also by taking the link costs into account when striving for feasibility.

6.2 Modelling extensions

A highly interesting subject for further research is the application of the augmented Lagrangean solution principle to other extended traffic assignment models than the one studied here. Such extended models may for example include side constraints describing some traffic control policy, limitations on traffic flows at intersections, joint capacities on two-way streets, requirements that observed flows on some links should be reproduced in the calculated solution, or dynamic aspects. From the results obtained in this work, we conjecture that such side constrained traffic assignment models can be efficiently dealt with computationally, although these constraints destroy the product structure. A confirmation of this conjecture through future research could hopefully lead to a renewed interest into the art of modelling real-world traffic problems. Important to note is that the application of the augmented Lagrangean solution principle to side constrained traffic assignment models provides a large flexibility in the design of the model, since this solution principle can handle both non-linear and non-separable side constraints.

As described at the end of Section 2.1, explicit link capacities may be used by the traffic engineer to calculate the appropriate corrections of tentative travel time functions. An interesting direction of future research is to develop and formalize this technique into a means for constructing travel time functions which take supplementary traffic flow restrictions into account. In such a procedure one would formulate and solve a traffic assignment problem (with relatively simple travel cost functions) which includes a set of suitable side constrains, and then utilize the optimal Lagrange multipliers for these constraints to derive adjusted travel time functions which indirectly take into account the additional model components. (Observe that the optimal values of the multipliers of course depend on the problem data, and therefore the adjusted travel time functions may not be valid for use in another problem instance.) This way of deriving improved descriptions of travel times may be to prefer to a calibration of parameters in non-standard travel cost functions, since it may be comparably easy to identify a set of appropriate side constraints and estimate the values of their coefficients, which may have very tangible physical interpretations.

The results of Section 2.1 are in Larsson and Patriksson (1995) extended to the case of general side constrained traffic assignment models.

Acknowledgements

We thank associate editor Michael G.H. Bell for encouragements and valuable suggestions. The research leading to this report was financially supported by the Swedish Research Council for Engineering Sciences (TFR) (Dnr. 92-824 and Dnr. 94-292) and the Swedish Transport and Communications Research Board (KFB) (Dnr. 15/88-62 and Dnr. 92-128-63). M.Sc. Håkan Fortell implemented the proposed solution method and made the

computational experiments within his final project.

References

[1] Aashtiani, H.Z. (1979). *The multi-modal traffic assignment problem.* Doctoral dissertation, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

[2] Aashtiani, H.Z. and Magnanti, T.L. (1981). Equilibria on a congested transportation network. *SIAM Journal on Algebraic and Discrete Methods, 2*, 213–226.

[3] Anantharamaiah, K.M. (1974). Equilibrium conditions in traffic assignment. In D.J. Buckley (Ed.), *Transportation and Traffic Theory, Proceedings of the Sixth International Symposium on Transportation and Traffic Theory*, Sydney, Australia (pp. 483–493). New York, NY: Elsevier.

[4] Barton, R.R. and Hearn, D.W. (1979). *Decomposition techniques for nonlinear cost multicommodity flow problems.* Research Report 79-2, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL.

[5] Bazaraa, M.S., Sherali, H.D. and Shetty, C.M. (1993). *Nonlinear Programming: Theory and Algorithms.* Second Edition. New York, NY: John Wiley & Sons.

[6] Beckmann, M.J. and Golob, T.F. (1974). Traveler decisions and traffic flows: a behavioral theory of network equilibrium. In D.J. Buckley (Ed.), *Transportation and Traffic Theory, Proceedings of the Sixth International Symposium on Transportation and Traffic Theory*, Sydney, Australia (pp. 453–482). New York, NY: Elsevier.

[7] Beckmann, M., McGuire, C.B. and Winsten, C.B. (1956). *Studies in the Economics of Transportation.* New Haven, CT: Yale University Press.

[8] Bertsekas, D.P. (1975). Combined primal-dual and penalty methods for constrained minimization. *SIAM Journal on Control, 13*, 521–544.

[9] Bertsekas, D.P. (1982). *Constrained Optimization and Lagrange Multiplier Methods.* New York, NY: Academic Press.

[10] Boyce, D.E., Janson, B.N. and Eash, R.W. (1981). The effect on equilibrium trip assignment of different link congestion functions. *Transportation Research, 15A*, 223–232.

[11] Branston, D. (1976). Link capacity functions: a review. *Transportation Research, 10*, 223–236.

[12] Bureau of Public Roads (1964). *Traffic Assignment Manual.* U.S. Department of Commerce, Washington, D.C.

[13] Conn, A.R., Gould, N.I.M. and Toint, Ph.L. (1991). A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis, 28*, 545–572.

[14] Dafermos, S.C. (1972). The traffic assignment problem for multiclass-user transportation networks. *Transportation Science, 6*, 73–87.

[15] Dafermos, S. (1980). Traffic equilibrium and variational inequalities. *Transportation Science, 14*, 42–54.

[16] Dafermos, S.C. and Sparrow, F.T. (1969). The traffic assignment problem for a general network. *Journal of Research of the National Bureau of Standards, 73B*, 91–118.

[17] Daganzo, C.F. (1977a). On the traffic assignment problem with flow dependent costs—I. *Transportation Research, 11*, 433–437.

[18] Daganzo, C.F. (1977b). On the traffic assignment problem with flow dependent costs—II. *Transportation Research, 11*, 439–441.

[19] Ferris, M.C. (1991). Finite termination of the proximal point algorithm. *Mathematical Programming, 50*, 359–366.

[20] Fiacco, A.V. and McCormick, G.P. (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* New York, NY: John Wiley & Sons.

[21] Hearn, D.W. (1980). *Bounding flows in traffic assignment models.* Research Report 80-4, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL

[22] Hearn, D.W. and Lawphongpanich, S. (1989). A dual ascent algorithm for traffic assignment problems. In *Dynamic Control and Flow Equilibrium, Proceedings of the Italy-U.S.A. Joint Seminar on Urban Traffic Networks*, Naples and Capri, Italy (pp. 35–53).

[23] Hearn, D.W., Lawphongpanich, S. and Nguyen, S. (1984). Convex programming formulations of the asymmetric traffic assignment problem. *Transportation Research, 18B*, 357–365.

[24] Hearn, D.W., Lawphongpanich, S. and Ventura, J.A. (1987). Restricted simplicial decomposition: computation and extensions. *Mathematical Programming Study, 31*, 99–118.

[25] Hearn, D.W. and Ribera, J. (1980). Bounded flow equilibrium problems by penalty methods. In *Proceedings of the 1980 IEEE International Conference on Circuits and Computers* (pp. 162–166).

[26] Hearn, D.W. and Ribera, J. (1981). Convergence of the Frank–Wolfe method for certain bounded variable traffic assignment problems. *Transportation Research, 15B*, 437–442.

[27] Hestenes, M.R. (1969). Multiplier and gradient methods. *Journal of Optimization Theory and Applications, 4*, 303–320.

[28] Hestenes, M.R. (1975). *Optimization Theory: The Finite Dimensional Case.* New York, NY: John Wiley & Sons.

[29] Inouye, H. (1987). Traffic equilibria and its solution in congested road networks. In R. Genser (Ed.), *Proceedings of the IFAC Conference on Control in Transportation Systems*, Vienna, 1986 (pp. 267–272).

[30] Jorgensen, N.O. (1963). *Some aspects of the urban traffic assignment problem.* Graduate Report, Institute of Transportation and Traffic Engineering, University of California, Berkeley, CA.

[31] Kennington, J.L. and Shalaby, M. (1977). An effective subgradient procedure for minimal cost multicommodity flow problems. *Management Science, 23*, 994–1004.

[32] Klessig, R.J. (1974). An algorithm for nonlinear multicommodity flow problems. *Networks, 4*, 343–355.

[33] Kort, B.W. and Bertsekas, D.P. (1976). Combined primal-dual and penalty methods for convex programming. *SIAM Journal on Control and Optimization, 14*, 268–294.

[34] Larsson, T. and Liu, Z.-W. (1989). *A Lagrangean relaxation scheme for structured linear programs with application to multicommodity network flows.* Report LiTH-MAT-R-89-24, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden. Revised 1992.

[35] Larsson, T., Liu, Z.-W. and Patriksson, M. (1992). *A dual scheme for traffic assignment problems.* Report LiTH-MAT-R-92-21, Department of Mathematics, Linköping Institute of Technology, Linköping, Sweden.

[36] Larsson, T. and Patriksson, M. (1992). Simplicial decomposition with disaggregated representation for the traffic assignment problem. *Transportation Science, 26*, 4–17.

[37] Larsson, T. and Patriksson, M. (1994). An augmented Lagrangean scheme for capacitated traffic assignment problems. In F. Boillot, N. Bhouri, and F. Leurent (Eds.), *Proceedings of the 2nd Meeting of the EURO Working Group on Urban Traffic and Transportation*, Paris, France, September 15–17, 1993 (pp. 163–199). Vol. 38 of Actes INRETS. Arcueil, France: Institut National de Recherche sur les Transports et leur Sécurité (INRETS).

[38] Larsson, T. and Patriksson, M. (1995). On side constrained traffic assignment models—equilibrium characterizations of solutions and an algorithm principle. *Transportation Research, forthcoming.*

[39] Lasdon, L.S. (1970). *Optimization Theory for Large Systems.* New York, NY: MacMillan.

[40] LeBlanc, L.J., Morlok, E.K. and Pierskalla, W.P. (1975). An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Science, 19*, 445–462.

[41] Luenberger, D.G. (1984). *Linear and Nonlinear Programming.* Second Edition. Reading, MA: Addison-Wesley.

[42] Miller, S.D., Payne, H.J. and Thompson, W.A. (1975). An algorithm for traffic assignment on capacity constrained transportation networks with queues. Paper presented at the *Johns Hopkins Conference on Information Sciences and Systems*, Johns Hopkins University, Baltimore, MD, April 2–4, 1975.

[43] Nagurney, A. (1993). *Network Economics: A Variational Inequality Approach.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

[44] Nguyen, S. (1976). A unified approach to equilibrium methods for traffic assignment. In M.A. Florian (Ed.), *Traffic Equilibrium Methods, Proceedings of the International Symposium in Montreal* (pp. 148–182). Lecture Notes in Economics and Mathematical Systems, Vol. 118. New York, NY: Springer-Verlag.

[45] Nguyen, S. and Dupuis, C. (1984). An efficient method for computing traffic equilibria in networks with asymmetric transportation costs. *Transportation Science, 18*, 185–202.

[46] Patriksson, M. (1994). *The Traffic Assignment Problem: Models and Methods.* Utrecht, The Netherlands: VSP.

[47] Payne, H.J. and Thompson, W.A. (1975). Traffic assignment on transportation networks with capacity constraints and queueing. Paper presented at the *47th National ORSA Meeting/TIMS 1975 North-American Meeting*, Chicago, IL, April 30–May 2, 1975.

[48] Pierre, D.A. and Lowe, M.J. (1975). *Mathematical Programming Via Augmented Lagrangians.* Reading, MA: Addison-Wesley.

[49] Pigou, A.C. (1920). *The Economics of Welfare.* London: MacMillan & Co.

[50] Polak, J. (1983). Some methodological aspects of equilibrium assignment algorithms. Paper presented at the *Annual Conference of the Universities' Transport Study Group.*

[51] Polyak, B.T. (1987). *Introduction to Optimization.* New York, NY: Optimization Software.

[52] Powell, M.J.D. (1969). A method for nonlinear constraints in optimization problems. In R. Fletcher (Ed.), *Optimization* (pp. 283–298). New York, NY: Academic Press.

[53] Rockafellar R.T. (1973a). A dual approach to solving nonlinear programming problems by unconstrained optimization. *Mathematical Programming, 5*, 354–373.

[54] Rockafellar, R.T. (1973b). The multiplier method of Hestenes and Powell applied to convex programming. *Journal of Optimization Theory and Applications, 12*, 555–562.

[55] Rockafellar, R.T. (1976a). Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research, 1*, 97–116.

[56] Rockafellar, R.T. (1976b). Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization, 14*, 877–898.

[57] Sender, J.G. and Netter, M. (1970). *Equilibre offre-demande et tarification sur un réseau de transport.* Institut de Recherche des Transport, Arcueil, France.

[58] Sheffi, Y. (1985). *Urban Transportation Networks: Equilibrium Analysis with Mathematical Methods.* Englewood Cliffs, NJ: Prentice-Hall.

[59] Smith, M.J. (1979). The existence, uniqueness and stability of traffic equilibria. *Transportation Research, 13B*, 295–304.

[60] Smith, M.J. (1987). Traffic control and traffic assignment in a signal-controlled network with queueing. Paper presented at the *Tenth International Symposium on Transportation and Traffic Theory*, Boston, MA.

[61] Stefek, D. (1989). *Extensions of simplicial decomposition for solving the multicommodity flow problem with bounded arc flows and convex costs.* Doctoral dissertation, University of Pennsylvania.

[62] Tomlin, J.A. (1966). Minimum-cost multicommodity network flows. *Operations Research, 14*, 45–51.

[63] Tseng, P. and Bertsekas, D.P. (1993). On the convergence of the exponential multiplier method for convex programming. *Mathematical Programming, 60*, 1–19.

[64] Vanderstraeten-Tilquin, G. (1977). *Problèmes de circulation avec coûts convexes.* Publication 268, Département d'Informatique et de recherche opérationelle, Université de Montréal, Montréal.

[65] von Hohenbalken, B. (1977). Simplicial decomposition in nonlinear programming algorithms. *Mathematical Programming, 13*, 49–68.

[66] Wardrop, J.G. (1952). Some theoretical aspects of road traffic research. In *Proceedings of the Institute of Civil Engineers*, Part II (pp. 325–378).

[67] Yang, H. and Yagar, S. (1994). Traffic assignment and traffic control in general freeway-arterial corridor systems. *Transportation Research, 28B*, 463–486.

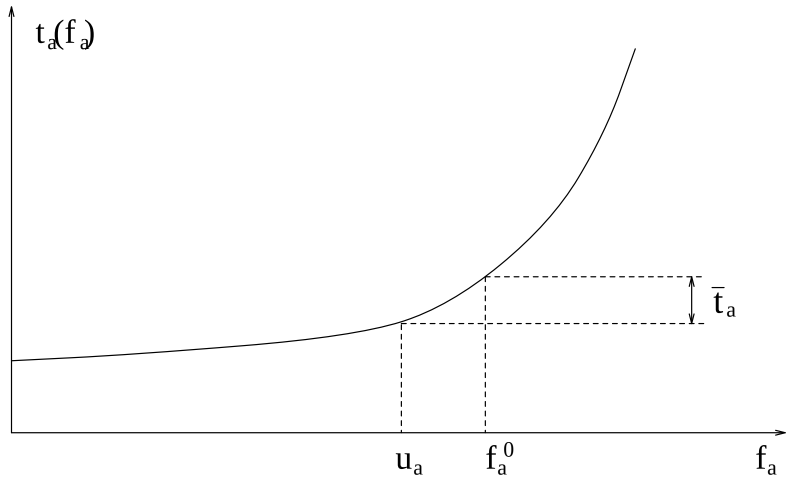

Figure 1: Finding initial values for the Lagrange multipliers

No.	City	Reference	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{C}\rvert$
1		Nguyen and Dupuis (1984)	9	13	4
2		Barton and Hearn (1979)	9	18	4
3		Hearn and Ribera (1980)	9	18	4
4	Sioux Falls	LeBlanc *et al.* (1975)	24	76	528
5	Winnipeg	Nguyen (1976)	1052	2836	4345

Table 1: Test networks

Link	Capacity	A	B
(1,5)	12	12.00	12.01
(1,6)	18	18.00	17.99
(2,5)	35	34.98	35.00
(2,6)	35	35.01	35.00
(5,6)	20	9.97	10.06
(5,7)	11	11.01	10.97
(5,9)	26	26.00	25.98
(6,5)	11	0.00	0.00
(6,8)	33	33.00	33.00
(6,9)	32	30.00	30.05
(7,3)	25	24.93	25.01
(7,4)	24	16.98	17.00
(7,8)	19	0.00	0.00
(8,3)	39	15.06	14.99
(8,4)	43	43.02	43.00
(8,7)	36	4.91	5.04
(9,7)	26	25.99	26.00
(9,8)	30	30.00	30.02

Table 2: Comparison of two augmented Lagrangean algorithms

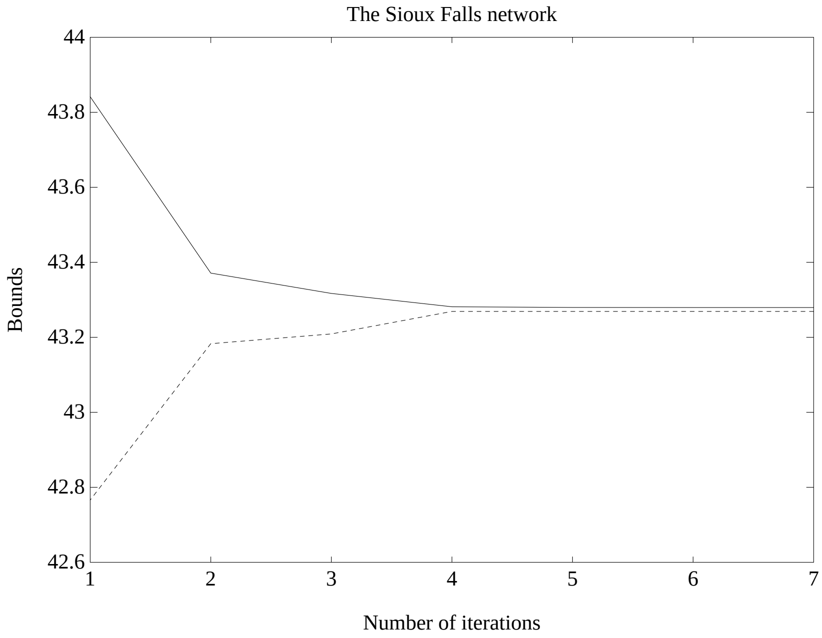

Figure 2: Upper and lower bounds versus iterations