An Augmented Lagrangean Scheme for Capacitated 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 4, 1995

Abstract

The inclusion of explicit bounds on the link flows in traffic assignment models has been proposed as a means to obtain a more accurate description of the traffic behaviour than that given by the basic, uncapacitated, model. Although the capacitated problem has the advantage of being numerically more tractable than models involving travel cost functions which tend to infinity as flows approach the link capacities, it has received very limited attention in the past. The main reason for this is that the explicit capacities make the problem computationally more demanding, since they destroy the Cartesian product structure of the feasible set of the uncapacitated model, which enables the development of highly efficient solution procedures for that problem.

The availability of efficient procedures for the basic model motivates the use of dualization approaches for handling the capacity constraints of the more complex model; we propose and evaluate an augmented Lagrangean method in which uncapacitated traffic assignment subproblems are solved with the disaggregate simplicial decomposition algorithm. This algorithm fully exploits the subproblems' structure and also has very favourable reoptimization facilities. Both these properties are of greatest importance for achieving computational efficiency of augmented Lagrangean schemes. The overall method exhibits a linear rate of convergence under standard nondegeneracy assumptions.

Experiments with medium-scale traffic networks with varying levels of capacities demonstrate that the proposed algorithm is efficient and solves capacitated problems within a factor of, roughly, three to four of the computing times needed to solve the corresponding uncapacitated problems; also, the augmented Lagrangean scheme compares very favourably with both ordinary Lagrangean dualization and penalty methods.

The suggested strategy can be utilized for the solution of other extensions of the basic traffic assignment problem where explicit constraints are added in order to refine the basic model; these constraints can be used to model, for instance, traffic flows at intersections, or different categories of vehicles. This way of extending the basic model may be an interesting alternative to the use of asymmetric traffic assignment models, where such extensions are made through the development of complex travel cost functions, which, in practical applications, are difficult to calibrate.

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

1 Introduction

A class of mathematical models which is frequently studied in the field of urban traffic planning is the so-called *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 or predicting a traffic flow pattern in a road network, where there is some (fixed or elastic) demand for travels 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 principle of optimality 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; it is therefore known as the principle of *user equilibrium*. Under the assumptions that the functions describing the travel times on the links of the road network are integrable and monotone, Wardrop's first principle of optimality gives rise to a convex mathematical program (e.g. Dafermos, 1972). First to formulate this program are Beckmann *et al.* (1956), in the case of separable travel cost functions. Travel cost functions which are non-integrable have asymmetric Jacobian matrices, and the resulting problems are therefore known as *asymmetric traffic assignment problems*. Such problems typically arise when modelling flows of different modes (e.g. Dafermos, 1972), or when the traffic flows on the links interact in the sense that the travel cost for one link depends on traffic flows on other links, and it may be formulated as, for instance, 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). The second optimality principle is known as the *system optimum* principle, and corresponds to the situation in which the total travel time is minimized. Under the assumption that the travel cost 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 basic model of traffic assignment has received a lot of attention and several highly efficient solution methods for it have been developed. (See e.g. Patriksson, 1990, for a review of models and methods for traffic assignment problems). One important reason for its popularity is that its large size in practical applications combined with its very special structure makes it a challenge for research regarding the development of efficient special-purpose solution methods. The reader should in particular recall that the Cartesian product structure which is inherent in the basic problem is exploited in all efficient algorithms for it (e.g. Larsson and Patriksson, 1992). Another reason for the popularity of the basic model is its simplicity and nice interpretations, which make it attractive for practical use.

However, its inherent simplicity makes it inapplicable to more complex real-world traffic problems (e.g. Sender and Netter, 1970). In order to improve a traffic assignment model's ability to accurately describe, reproduce, or predict a real-world traffic situation, one may consider two fundamentally different modelling strategies.

A first alternative is to capture complex flow relationships through the introduction of, 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 the references cited therein). Seemingly, the asymmetric models' popularity is a consequence of their mathematical elegance and nice interpretations rather than their applicability, since real-world applications seem to be lacking. A major reason for this is probably the practical difficulty of calibrating the asymmetric travel time functions.

A second possibility to improve the quality of a traffic assignment model is to capture complex flow relationships and supplementary flow restrictions through the introduction of an additional set of constraints. Such constraints may for example describe joint capacities on two-way links, limitations on total flows through junctions, restrictions arising from traffic control policies, or dynamic aspects. We believe this second approach to be much more appealing from a practical point of view, since it is certainly easier for the traffic engineer to identify a set of suitable side constraints (which may have immediate physical interpretations), than to make proper estimates of parameter values in complex travel time functions.

Although the second alternative seems to be the most useful from a practical point of view, it has been given a very limited attention compared to the first one. As a starting point for further work along the lines of the second alternative, we present an efficient solution approach for the capacitated traffic assignment problem. The solution principle can be applied to general side constrained traffic assignment models, and the results presented in this report indicate that the inclusion of side constraints does not necessarily make the model computationally intractable.

There have been many suggestions for choices of classes of functions to be used to model the congestion effects, i.e., to describe the dependencies between traffic flows and travel times (e.g. Branston, 1976). In practice, the most frequently used functions are polynomials whose degrees and coefficients are determined from real-world data through statistical methods. These travel time functions are however not realistic in the sense that the resulting travel times will become finite whenever the link flows are finite, so that the links are actually assumed to be able to carry arbitrarily large volumes of traffic flows although the links will in practice undoubtedly have some finite limits on traffic flows, because of, for instance, congestion, speed limit regulations, or cycle-times for traffic-signals. 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." An obvious and very simple way of improving the quality of a traffic assignment model would thus be to include link flow capacities. This can be done either explicitly through the introduction of explicit upper bounds on link flows, or implicitly through the use of travel time functions which tend to infinity when the link flows approach their respective capacities (Daganzo, 1977a and 1977b), but neither of these two techniques have been studied to any greater extent.

From a modelling point of view, explicit upper bounds have the advantage of allowing the link flows to attain the capacity values, whereas the use of travel time functions with an asymptote at the capacity level will force all link flows to be strictly less than the capacities. A disadvantage of imposing explicit capacities is that they destroy the profitable Cartesian product structure which is inherent in the uncapacitated problem. For example, 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 prohibitively computationally expensive in practice. (The loss of the product structure in the feasible set is without doubt the main reason why traffic assignment problems with explicit capacities have been so little studied.) 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 replaced by shortest-route subproblems while maintaining convergence to an optimal flow pattern (Daganzo, 1977a and 1977b, and Hearn and Ribera, 1981). (Note that a shortest-route subproblem is obtained if the coupling capacity constraints of the linear multicommodity flow subproblem are relaxed.) Further, a solution to an explicitly capacitated traffic assignment problem will, in the user equilibrium case, no longer comply with Wardrop's first principle (Hearn, 1980), which has over the years been established and accepted as the fundamental behavioural assumption; however, the solution will satisfy Wardrop's first principle if the usual travel costs are replaced by certain well-defined *generalized* travel costs. From a computational point of view, 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 (such as 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) capacities (Daganzo, 1977b); this task is non-trivial though (see e.g. Inouye, 1986).

Because of the availability of several highly efficient solution methods for the uncapacitated traffic assignment problem, a very natural strategy when designing a solution method for the capacitated problem 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 problem are therefore based on penalization and/or dualization concepts, see Jorgensen (1963), Vanderstraeten-Tilquin (1977), Hearn (1980), Hearn and Ribera (1980), Polak (1983), and Inouye (1986). However, in a recent Ph.D. thesis, Stefek (1989) proposes and studies extensions of the Frank-Wolfe and simplicial decomposition methods for the uncapacitated problem (see e.g. LeBlanc *et al.,* 1975, and Hearn *et al.,* 1987, respectively) to the capacitated problem.

The capacitated problem constitutes a special case of the general side constrained traffic assignment problem, and the main objective with the research to be presented here is to show that such problems may be efficiently solved, although the side constraints destroy the Cartesian product structure of the basic problem; this is accomplished through a combination of an augmented Lagrangean dualization scheme (see e.g. Bertsekas, 1982, for a general description of this class of methods) and an efficient solution method for the basic traffic assignment problem, the disaggregate simplicial decomposition (DSD) method (Larsson and Patriksson, 1992).

Augmented Lagrangean dual schemes can be derived as combinations of ordinary Lagrangean dual schemes and traditional penalty methods, but have the advantages that the multiplier values in general converge to the optimal ones much faster than in ordinary Lagrangean dual schemes, and that the numerical ill-conditioning that is inherent in traditional penalty methods is avoided. The main difference between the augmented Lagrangean scheme to be proposed below as compared to the ones previously studied for the capacitated traffic assignment problem is that we employ a much more efficient algorithm for the solution of the uncapacitated subproblems. The DSD algorithm fully exploits the underlying problem structure, and, in addition, has very good reoptimization properties, which is of outmost importance in the application under consideration.

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

The resulting overall solution method for the capacitated traffic assignment problem is applied to some test problems of varying sizes; these problems are derived from well-known uncapacitated problems through the introduction of properly chosen explicit link capacities. Our computational experience is that the capacitated problems are solved within computing times that are no more than about three to four times longer than the computing times for the corresponding uncapacitated problems. We have thus shown that the traffic assignment model may be computationally tractable even though it is extended through the addition of side constraints; this is the main contribution with this work. It is our hope that this conclusion may lead to a renewed interest into the question of how to formulate proper mathematical models of real-world traffic assignment problems.

The outline of the remainder of the report is as follows. In Section 2, we introduce the mathematical model for the capacitated user equilibrium traffic assignment problem, state its optimality conditions and review the previously suggested solution methods. We next establish the overall augmented Lagrangean scheme together with two convergence results. The fourth section describes in detail the realizations of the steps of the overall scheme, and Section 5 gives computational results for networks of varying sizes, aiming at illustrating various characteristics of the algorithm and proving its viability. We finalize by drawing some conclusions and suggesting some interesting 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 arcs, respectively. (In mathematical contexts, the links are sometimes referred to as arcs.) For certain ordered pairs of nodes, $(p, q) \in \mathcal{C}$, where node p is the origin, node q the destination, and $\mathcal{C} \subset \mathcal{N} \times \mathcal{N}$, there are given positive flow demands d_{pq}, $(p,q) \in \mathcal{C}$, which give rise to a link flow pattern $\mathbf{f}$ when distributed through the network. Associated with each directed arc $a \in \mathcal{A}$ is a positive transportation cost function, $t_a(f_a)$, which measures the disutility of using the arc as a function of the network flow. As a result of congestion, the cost functions, or *link performance functions*, are assumed to be monotonically increasing. We will consider the problem of determining a user equilibrium traffic flow pattern fulfilling the travel demands.

2.1 The Wardrop conditions

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 network flow 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 travel time on the shortest-route from p and q, the Wardrop user equilibrium conditions may equivalently be stated as

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

to hold for all pairs $(p,q) \in \mathcal{C}$, so that the optimality conditions for the user equilibrium traffic assignment problem may be summarized as

$$\begin{array}{rclll} c_{pqr} & \geq & \pi_{pq} & \forall r \in \mathcal{R}_{pq} & \forall (p,q) \in \mathcal{C} \\ \sum_{r \in \mathcal{R}_{pq}} h_{pqr} & = & d_{pq} & & \forall (p,q) \in \mathcal{C} \\ h_{pqr} & \geq & 0 & \forall r \in \mathcal{R}_{pq} & \forall (p,q) \in \mathcal{C} \\ h_{pqr}(c_{pqr} - \pi_{pq}) & = & 0 & \forall r \in \mathcal{R}_{pq} & \forall (p,q) \in \mathcal{C}. \end{array}$$

Here, the second and third sets of conditions ensure feasibility with respect to the prescribed origin-destination travel demands and non-negativity of route flows, respectively.

As was established by Beckmann *et al.* (1956), one may derive a convex mathematical program whose Karush-Kuhn-Tucker conditions coincide with the above set of conditions; further, since the Guignard constraint qualification is always fulfilled for linearly constrained programs, these conditions turn out to be both necessary and sufficient.

2.2 The capacitated model and its optimality conditions

Letting u_a, $a \in \mathcal{A}$, denote positive explicit upper bounds on the arc flows, and introducing the arc-route incidence matrix (δ_{pqra}) defined by

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

the arc-route formulation (see e.g. Dafermos and Sparrow, 1969) of the capacitated user equilibrium traffic assignment problem is

$$\begin{array}{lllllll}
[\textbf{TAP-C}] & \min & T(\mathbf{f}) = \sum_{a\in\mathcal{A}} \int_0^{f_a} t_a(s)ds & & & & \\
& \text{s.t.} & \sum_{r\in\mathcal{R}_{pq}} h_{pqr} & = & d_{pq} & \forall (p,q) \in \mathcal{C} & (1) \\
& & h_{pqr} & \geq & 0 \quad \forall r \in \mathcal{R}_{pq} & \forall (p,q) \in \mathcal{C} & (2) \\
& & \sum_{(p,q)\in\mathcal{C}} \sum_{r\in\mathcal{R}_{pq}} \delta_{pqra} h_{pqr} & = & f_a \quad \forall a \in \mathcal{A} & & (3) \\
& & f_a & \leq & u_a \quad \forall a \in \mathcal{A}. & & (4)
\end{array}$$

Disregarding the constraints (4), we obtain the basic model of traffic assignment, which will be referred to as [**TAP**]. Furthermore, the inherent Cartesian product structure of that model is recovered by eliminating its (auxiliary) link flow variables. We will throughout the paper presume that [**TAP-C**] has a feasible solution, i.e., that the explicit link capacities are large enough to allow all demands for travels to be distributed through the network.

The problem [**TAP-C**] is a convex mathematical program, with optimal solutions which are unique with respect to link flows because of the strict convexity of the objective. (There are, in general, alternative route-flow solutions.) Letting ψ_{pq}, $(p,q) \in \mathcal{C}$, and β_a, $a \in \mathcal{A}$, denote the optimal values of the Lagrangean multipliers for the demand feasibility constraints (1) and the capacity constraints (4), respectively, the necessary and sufficient optimality conditions for [**TAP-C**] may be stated as

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

Similar optimality conditions have been given by Jorgensen (1963), Hearn (1980), and Inouye (1986).

Introducing a generalized route cost

$$\overline{c}_{pqr} = c_{pqr} + \sum_{a \in \mathcal{A}} \delta_{pqra} \beta_a,$$

a capacitated user equilibrium flow may equivalently be characterized by the conditions

$$\begin{array}{rclcrcl} h_{pqr} & > & 0 & \Rightarrow & \overline{c}_{pqr} & = & \psi_{pq} \\ h_{pqr} & = & 0 & \Rightarrow & \overline{c}_{pqr} & \geq & \psi_{pq}, \end{array}$$

i.e., the multiplers ψ_{pq}, $(p, q) \in \mathcal{C}$, may be interpreted as the generalized travel costs of the utilized routes and, further, the generalized travel cost on any non-utilized route in an origin-destination pair is at least as large as the generalized travel cost on the routes utilized in the pair. An equilibrium flow may thus still be described by Wardrop-type conditions with actual travel costs replaced by generalized travel costs.

Further, assuming, without any loss of generality, that the first l routes are actually used, and that among these the first m are *saturated*, i.e., contain at least one link which carries flow on its capacity level, the network flow is a capacitated user equilibrium if and only if

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

(cf. Jorgensen, 1963, for the case of constant travel times), and the unused routes in the origin-destination pair have generalized route costs that are at least as large as that of the used routes, for any pair $(p, q) \in \mathcal{C}$.

It is here important to note that one can, in general, not relate the actual travel costs of the unused routes to those of the used ones; it may for instance happen that the cheapest route in an origin-destination pair is not used because its generalized cost becomes too high. Furthermore, one can for the capacitated problem not formulate a simple optimality condition, similar to Wardrop's first principle, in terms of actual travel costs. (The extensions of Wardrop's first principle stated by Anantharamaiah, 1974, and Stefek, 1989, are incorrect or, possibly, poorly formulated.) This is due to the fact that the Wardrop principles are intimately associated with the Cartesian product structure of the feasible set of [**TAP**].

The Lagrangean multipliers for the capacity constraints, β_a, $a \in \mathcal{A}$, "measure the time gained by users of routes filled to capacity compared to the fastest route still available" (Jorgensen, 1963), but can also be given other interesting interpretations. First, in a network where some links have queues at their exits (i.e., where some links are saturated), they may be interpreted as the equilibrium time delays caused by the queuing (Payne and Thompson, 1975; see also Smith, 1987); this result is extended to the case of non-constant link travel times by Miller *et al.* (1975) and Inouye (1986). Second, they can be seen as 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. (One may note that although the equilibrium generalized route travel costs ψ_{pq}, $(p, q) \in \mathcal{C}$ are unique, this is not necessarily true for the multipliers β_a, $a \in \mathcal{A}$.) Beckmann and Golob (1974) make a similar observation for a capacitated, *system* optimum assignment model, where the multipliers are interpreted as link tolls which achieve a system optimum when the individual travellers minimize their respective generalized travel costs.

It is important to note that the capacitated equilibrium link flow pattern found by solving [**TAP-C**] may also be found by solving the corresponding uncapacitated problem [**TAP**] with travel time functions adjusted to $t_a(f_a) + \beta_a$, $\forall a \in \mathcal{A}$. The solution of a capacitated problem can therefore be used as a tool 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. Furthermore, 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.

2.3 Review of previous solution methods

As has been previously mentioned, the explicit capacities introduced in [**TAP-C**] destroy the Cartesian product structure of the feasible set of [**TAP**]. In an application of the Frank-Wolfe algorithm to [**TAP-C**], subproblems do therefore not separate into a number of independent shortest-route problems, but instead define linear multicommodity problems (Klessig, 1974). Algorithms proposed for the solution of [**TAP-C**] are often based on the recognition that such subproblems are prohibitively expensive to solve repeatedly. These methods 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 capacity constraints and, to ensure convergence, the cost functions t_a are assumed to satisfy the *coercivity* condition

$$\lim_{f_a \to u_a} \int_0^{f_a} t_a(s)ds = +\infty, \ \forall a \in \mathcal{A}, \tag{2.1}$$

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, but the most natural assumption on the travel time functions which implies this sufficient condition is, however, Condition (2.1). 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 Stefek (1989).

The main theme of Stefek's dissertation is however 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 multidimensional 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 in the simplicial decomposition methods also; whenever the extreme points corresponding to the shortest-route patterns do not provide sufficient descent in the master problem, a linear multicommodity flow subproblem is invoked. (Note that in a direct application of simplicial decomposition, subproblems would always be linear multicommodity flow problems, and the capacity constraints would not be invoked in the master problem; such a scheme would, however, not be efficient, because of the high computational cost of the subproblem.) Stefek also presents a variation in which Lagrangean 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 linear multicommodity flow subproblem have to be invoked. Computational experiments with three medium- and large-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 linear multicommodity flow subproblems are solved by a Dantzig-Wolfe decomposition approach), but inferior for heavily capacitated ones.

In the second approach, the capacitated problem is converted into a sequence of uncapacitated problems through a dualization/penalization of the capacity constraints (4). By using such a strategy, efficient methods for [**TAP**] may be applied for the solution of [**TAP-C**]. (There exists methods for [**TAP-C**] based on other relaxation strategies; in Hearn and Lawphongpanich, 1989, and Larsson *et al.,* 1992, the definitional constraints (3) are instead dualized, and in Inouye (1986), all constraints but (3) are dualized in an augmented Lagrangean fashion.)

For the case of constant travel times, Jorgensen (1963) suggests applying the Dantzig-Wolfe decomposition method (a column generation technique, which may be interpreted as a dual cutting plane method), but does not give any computational results. For the case of flow-dependent travel times, he suggests making an approximation through the use of piece-wise constant travel time functions; the approximate problem may then be restated and solved as a problem with constant travel times, in an enlarged network. Miller *et al.* (1975) also present a column generation approach for the flow-independent travel cost case, where 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 manner, whereas in the scheme of Hearn and Ribera, the subproblems are solved by the Frank-Wolfe method. Hearn and Ribera consider two different augmented Lagrangean functions and applies 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, which 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 draws the conclusion 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.

3 The augmented Lagrangean scheme

We introduce the set

$$\mathbf{F} = \left\{ \mathbf{f} \in \Re^{|\mathcal{A}|} \;\middle|\; \exists \mathbf{h} \text{ such that (1)–(3) holds} \right\},$$

i.e., the enlarged set obtained when ignoring the capacity constraints in the feasible set of [**TAP-C**].

In an *exterior penalty method*, the relaxed constraints (4) are included in the objective as a nonlinear function of the constraint functions

$$g_a(f_a) \stackrel{\text{def}}{=} f_a - u_a, \;\; a \in \mathcal{A},$$

such that the objective is penalized whenever a constraint in (4) is violated (i.e., whenever $g_a(f_a) > 0$ for some a).

We introduce a penalty function $P : \Re^{|\mathcal{A}|} \mapsto \Re$, satisfying

(1) $P(\mathbf{f}) \geq 0, \forall \mathbf{f} \in \mathbf{F}$,

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

(3) P is continuous on $\mathbf{F}$.

One possible choice of penalty function P satisfying **(1)**–**(3)** is

$$P(\mathbf{f}) = \sum_{a \in \mathcal{A}} p_a(f_a), \tag{3.1}$$

where

$$p_a(f_a) = r_a [g_a(f_a)]_+^{m_a} \stackrel{\text{def}}{=} r_a \max\{0, g_a(f_a)\}^{m_a}, \;\; r_a > 0, \;\; m_a \geq 2. \tag{3.2}$$

This penalty function is used by Hearn (1980).

The properties of penalty function methods are well known; see for example Fiacco and McCormick (1968). Let

$$\mathbf{f}(c) = \arg\min_{\mathbf{f}\in\mathbf{F}} L_c(\mathbf{f}) = T(\mathbf{f}) + cP(\mathbf{f})$$

be the solution of the penalty subproblem for a certain value of the *penalty parameter* $c > 0$, and

$$L_c = \min_{\mathbf{f}\in\mathbf{F}} L_c(\mathbf{f}).$$

We then have:

(1) $L_c \leq T(\mathbf{f}^*)$, $\forall c > 0$,

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

Furthermore, with a choice of a differentiable and separable penalty function P, such as the choice (3.1), (3.2), it can be shown (Hearn, 1980) that

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

that is, an estimate of optimal values of the Lagrange multipliers for the constraints (4) is available. (Note that this estimate may be poor, since it is based on two quantities, one with limit infinity and the other with limit zero.)

In order to avoid the instabilities and ill-conditioning inherent in the penalty approach, we introduce a Lagrangean term in the penalized objective, thus creating an *augmented Lagrangean function* (Hestenes, 1969, Powell, 1969, Rockafellar, 1973b, and Bertsekas, 1975 and 1982).

If the penalty function is of the form (3.2), with $r_a = \frac{1}{2}$ and $m_a = 2$, $\forall a \in \mathcal{A}$, the augmented Lagrangean becomes (Rockafellar, 1973a)

$$L_c(\mathbf{f}, \boldsymbol{\mu}) = T(\mathbf{f}) + \sum_{a\in\mathcal{A}} \overline{p}_a(f_a, \mu_a, c), \tag{3.3}$$

where $\boldsymbol{\mu}$ is the vector of Lagrangean multipliers for the constraints (4), and

$$\overline{p}_a(f_a, \mu_a, c) = \frac{1}{2c}([\mu_a + cg_a(f_a)]_+^2 - \mu_a^2). \tag{3.4}$$

Introducing

$$\begin{aligned} \mathbf{f}(\boldsymbol{\mu}, c) &= \arg\min_{\mathbf{f}\in\mathbf{F}} L_c(\mathbf{f}, \boldsymbol{\mu}), \\ L_c(\boldsymbol{\mu}) &= \min_{\mathbf{f}\in\mathbf{F}} L_c(\mathbf{f}, \boldsymbol{\mu}), \end{aligned} \tag{3.5}$$

we have that

$$\mathbf{f}(\boldsymbol{\beta}, c) = \mathbf{f}^*, \ \forall c \geq 0. \tag{3.6}$$

In an augmented Lagrangean approach, there is thus no need to let the parameter c tend to infinity in order to obtain the optimal solution. (The fact that also $c = 0$ is feasible in (3.6) is due to the strict convexity of T; see Bazaraa and Shetty, 1979, Theorem 6.5.2.)

In order to use (3.6) to obtain a near-optimal solution to [**TAP-C**], good estimates of the optimal multiplier values $\boldsymbol{\beta}$ must be found. These multipliers may be found by solving the augmented Lagrangean dual problem

$$\max_{\boldsymbol{\mu}} L_c(\boldsymbol{\mu}). \tag{3.7}$$

Note that (3.7) reduces to the Lagrangean dual problem when $c = 0$. However, ordinary Lagrangean dualization is inferior to augmented Lagrangean schemes, because of its slow rate of convergence, and, from now on, we will hence assume that $c > 0$.

Some basic properties of the augmented Lagrangean dual function are given below (see Rockafellar, 1973a).

(1) $L_c(\boldsymbol{\mu}) \leq T(\mathbf{f}^*)$, $\forall \boldsymbol{\mu} \geq \mathbf{0}$,

(2) $L_c(\boldsymbol{\beta}) = \max_{\boldsymbol{\mu}} L_c(\boldsymbol{\mu}) = T(\mathbf{f}^*)$,

(3) L_c is concave and differentiable, with

$$\frac{\partial L_c(\boldsymbol{\mu})}{\partial \mu_a} = \max\{g_a(f_a(\boldsymbol{\mu}, c)), -\frac{\mu_a}{c}\}, \ \forall a \in \mathcal{A}. \tag{3.8}$$

The expression (3.8) gives a guideline as how to update a trial point $\boldsymbol{\mu}$. A step c from $\boldsymbol{\mu}$ in the direction $\nabla L_c(\boldsymbol{\mu})$ yields (see Bertsekas, 1982, p. 162)

$$\mu_a := [\mu_a + c g_a(f_a(\boldsymbol{\mu}, c))]_+, \ \forall a \in \mathcal{A}; \tag{3.9}$$

if c is sufficiently small, this step actually is an ascent step for L_c. Furthermore, if $\boldsymbol{\mu}$ is sufficiently close to an optimal dual solution $\boldsymbol{\beta}$, (3.9) defines an ascent step also for the Lagrangean dual function L_0; see Bertsekas (1982, Proposition 5.8). Note that for ordinary Lagrangean duality, other updating formulae than (3.9) must be used (see, e.g., Polyak, 1969); in this case, ascent in the Lagrangean dual function is not guaranteed.

While convergence is ensured for any choice of positive parameter c, a good practical performance demands for a careful selection of its size. For instance, there is a trade-off between a high theoretical dual convergence rate and the amount of ill-conditioning of the primal subproblems (3.5); see Luenberger (1984, Chapter 13). The proper choice of parameter must therefore be determined through experimentation; see Hestenes (1975) and Bertsekas (1982). A usual approach is to start with a low value of c, and increase it whenever a measure of the total infeasibility in the relaxed constraints has not improved sufficiently within a fixed number of iterations; see, e.g., Powell (1969). We thus introduce a sequence $\{c_k\}$ of penalty parameters, satisfying $c_{k+1} \geq c_k \geq \ldots \geq c_1 > 0$, and the primal-dual sequence $\{(\mathbf{f}^k, \boldsymbol{\mu}^k)\}$ in the augmented Lagrangean scheme is given by an initial guess $\boldsymbol{\mu}^1$ and

$$
\begin{aligned}
\mathbf{f}^k &= \mathbf{f}(\boldsymbol{\mu}^k, c_k), & (3.10)\\
\mu_a^{k+1} &= [\mu_a^k + c_k g_a(f_a^k)]_+, \;\; a \in \mathcal{A}, k = 1, 2, \ldots. & (3.11)
\end{aligned}
$$

We next give the convergence results for the scheme (3.10), (3.11), as applied to [**TAP-C**].

Theorem 3.1 *Let $\boldsymbol{\mu}^1$ be arbitrary, $\{(\mathbf{f}^k, \boldsymbol{\mu}^k)\}$ be given by (3.10), (3.11), and $\{c_k\}$ satisfy $c_k \geq c > 0$, $\forall k$. Then, $\{\mathbf{f}^k\} \to \mathbf{f}^*$ and every accumulation point of $\{\boldsymbol{\mu}^k\}$ (at least one exists) is a vector of optimal Lagrange multipliers for (4).*

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

Remark 3.1 If $c_k \equiv c > 0$, then the sequence $\{\boldsymbol{\mu}^k\}$ converges; furthermore, the subproblem (3.10) may be solved inexactly (see Rockafellar, 1973b, 1976a).

Theorem 3.1 may be strengthened to yield rate of convergence results. To be able to obtain such results, a strict complementarity condition must be imposed.

Assumption A (Strict complementarity) Let $(\mathbf{f}^*, \boldsymbol{\beta})$ be an optimal primal-dual pair and $\overline{\mathcal{A}} = \{a \in \mathcal{A} \mid f_a^* = u_a\}$. If $a \in \overline{\mathcal{A}}$, then $\beta_a > 0$.

This strict complementarity assumption amounts to the users of a saturated link being faced with an additional cost (e.g., delay in a queue, see Section 2.2).

Theorem 3.2 *Let $t_a : \Re \mapsto \Re$ be continuously differentiable on $\Re_+$, $\forall a \in \mathcal{A}$, and Assumption A hold. Let $\boldsymbol{\mu}^1$ be arbitrary, $\{(\mathbf{f}^k, \boldsymbol{\mu}^k)\}$ be given by (3.10), (3.11), and $\{c_k\}$ satisfy $c_{k+1} \geq c_k \geq c > 0$, $\forall k$.*

(a) *$\{(\mathbf{f}^k, \boldsymbol{\mu}^k)\} \to (\mathbf{f}^*, \boldsymbol{\beta})$. Furthermore, the convergence of $\{\mu_a^k\}$, $a \notin \overline{\mathcal{A}}$, is finite.*

(b) *If further $\limsup_{k\to\infty} c_k = c^* < +\infty$ and $\boldsymbol{\mu}^k \neq \boldsymbol{\beta}$, $\forall k$, then $\{\boldsymbol{\mu}^k\}$ converges linearly.*

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

Proof. Under the given assumptions on [**TAP-C**], Assumption (S^+) of Bertsekas (1982, p. 161) holds. But then, from Proposition 3.2 of the same reference, Assumption (S) of Bertsekas (1982, p. 104) holds.

(a) Follows from Theorem 3.1 and Bertsekas (1982, p. 162).

(b) Follows from Bertsekas (1982, Proposition 2.7) and the above.

(c) Follows from Bertsekas (1982, Proposition 2.7) and the above. □

Remark 3.2 Corresponding results for inexact solutions of (3.10) can be found in Kort and Bertsekas (1976), Bertsekas (1982) and Rockafellar (1976a).

Remark 3.3 Rockafellar (1976a and 1976b) shows that finite convergence is achieved for the augmented Lagrangean algorithm when the objective exhibits a sharpness property (Polyak, 1987, p. 136). It is known (Ferris, 1991) that finite convergence is achieved in the dually equivalent proximal point algorithm under the slightly less restrictive weak sharpness assumption, and hence the augmented Lagrangean algorithm is finite whenever (3.7) has a weak sharp solution. Sufficient conditions under which this property holds is a subject for further research.

Until now, nothing has been said about the implementational issues of the above augmented Lagrangean scheme. In the next section we shall describe how $\boldsymbol{\mu}^1$ and $\{c_k\}$ are chosen, how the subproblems (3.10) are solved, and how feasible solutions can be constructed from $\mathbf{f}^k$.

4 A realization of the augmented Lagrangean scheme

We first give the steps of the algorithm. Then, each step is described in detail.

Step 0 (*Initialization*) Choose $\boldsymbol{\mu}^1$ and c_1. Let $k = 1$.

Step 1 (*Subproblem solution*) Solve the augmented Lagrangean subproblem (3.10). The solution is $\mathbf{f}^k$.

Step 2 (*Feasible solution generation*) Generate a feasible solution from $\mathbf{f}^k$.

Step 3 (*Convergence check*) If a termination criterion holds → Stop. Otherwise, continue.

Step 4 (*Multiplier and penalty update*) Update $\boldsymbol{\mu}^k$ according to (3.11) and c_k. Let $k := k+1$, and go to Step 1.

4.1 Initialization of the Lagrangean multipliers

The initial choice of multipliers is a critical point in devising an efficient implementation, see, e.g., Bertsekas (1982, Section 2.2.5), and any *a priori* knowledge of the problem must be utilized to select $\boldsymbol{\mu}^1$ as close as possible to $\boldsymbol{\beta}$.

Let $\mathbf{f}^*$ be the solution of [**TAP**]. If $f_a^* \leq u_a$ for all a, then $\mathbf{f}^*$ solves [**TAP-C**], and $\boldsymbol{\beta} = \mathbf{0}$ is an optimal set of multipliers for the capacity constraints. Assume therefore that there is a nonempty subset $\overline{\mathcal{A}} \subset \mathcal{A}$ of arcs with $f_a^* > u_a$, $a \in \overline{\mathcal{A}}$. Since $\mathbf{f}^*$ solves [**TAP**], Wardrop's first conditions of equal travel times are satisfied. Consider constructing a flow $\tilde{\mathbf{f}}$ through

$$\tilde{f}_a = \begin{cases} u_a & \text{if} \quad a \in \overline{\mathcal{A}} \\ f_a^* & \text{if} \quad a \notin \overline{\mathcal{A}}. \end{cases}$$

This flow satisfies the capacity constraints. Furthermore, if we let

$$\mu_a^1 = \begin{cases} t_a(f_a^*) - t_a(u_a) & \text{if} \quad a \in \overline{\mathcal{A}} \\ 0 & \text{if} \quad a \notin \overline{\mathcal{A}}, \end{cases} \tag{4.1}$$

i.e., we add the cost $\bar{t}_a = t_a(f_a^*) - t_a(u_a)$ to the travel time of the links that are subject to a flow reduction (see Figure 4.1), then $(\boldsymbol{\mu}^1, \tilde{\mathbf{f}})$ satisfies the generalized Wardrop conditions, but $\tilde{\mathbf{f}}$ violates the flow conservation constraints because of the removal of flows from the links in $\overline{\mathcal{A}}$. If, however, these violations are small, the vector $\boldsymbol{\mu}$ given by (4.1) will be near-optimal. We therefore initialize $\boldsymbol{\mu}^1$ according to (4.1).

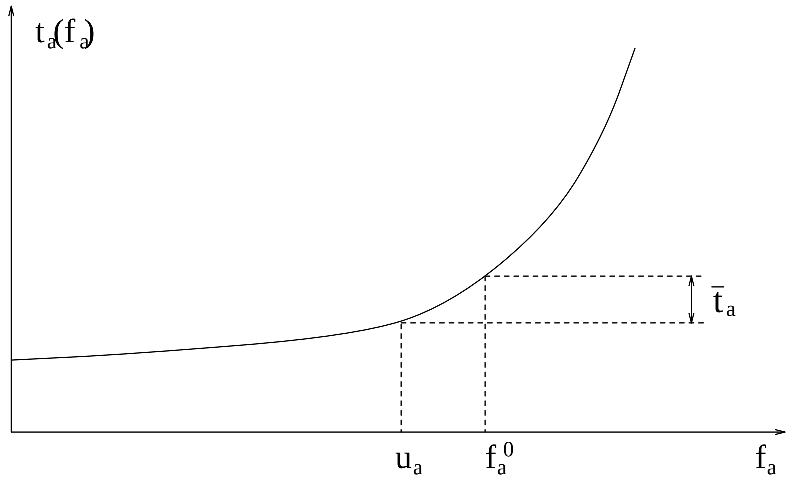

Figure 4.1: Finding start values for the Lagrange multipliers

4.2 Solving the subproblems

The subproblem (3.10), obtained when dualizing (4), is of the form [**TAP**], and may hence be solved with any of the existing methods for that problem. Since, however, we will solve a sequence of problems (3.10) with slightly varying data, we are interested in using a method with very good reoptimization facilities, that is, a method which can effectively utilize the solution of a previous subproblem when solving the next one. An algorithm with particularly nice reoptimization facilities is the disaggregate simplicial decomposition (DSD) algorithm (Larsson and Patriksson, 1992), which is briefly described below.

Consider the problem [**TAP**]. Disregarding the implicit link-flow defining constraints, the entire feasible set is actually a Cartesian product with respect to the separate commodities, that is, the constraint matrix has a block-diagonal shape. This enables the utilization of a simplicial decomposition algorithm using one convexity constraint in a master problem for each origin-destination pair.

In order to formulate the disaggregate master problem, we assume that a nonempty subset $\hat{\mathcal{R}}_{pq}$ of the set of simple routes $\mathcal{R}_{pq}$ is known. In the disaggregate master problem, each convexity variable λ_{pqi} will be associated with a route i in $\hat{\mathcal{R}}_{pq}$, and denote the portion of the demand d_{pq} distributed along route i. Let $\boldsymbol{\lambda}_{pq}$ be the vector of convexity variables λ_{pqi} for all $i \in \hat{\mathcal{R}}_{pq}$. The vector $\boldsymbol{\lambda}$ then consists of all such $\boldsymbol{\lambda}_{pq}$.

The link flows may be calculated from the weighted route flows according to:

$$f_a = \sum_{(p,q)\in\mathcal{C}} d_{pq} \sum_{i\in\hat{\mathcal{R}}_{pq}} \delta_{pqia}\lambda_{pqi}. \tag{4.2}$$

The function $T(\mathbf{f})$ is, through (4.2), a function of the convexity variables $\boldsymbol{\lambda}$; the notation $T(\boldsymbol{\lambda})$ will therefore be used in the following.

Utilizing the disaggregated representation the restricted master problem becomes:

$$[\mathbf{RMP}] \qquad \begin{array}{lrcll} \min & T(\boldsymbol{\lambda}) & & & \\ \text{s.t.} & \displaystyle\sum_{i\in\hat{\mathcal{R}}_{pq}} \lambda_{pqi} & = & 1 & \forall (p,q)\in\mathcal{C} \\ & \lambda_{pqi} & \geq & 0 \quad \forall i \in \hat{\mathcal{R}}_{pq} & \forall (p,q)\in\mathcal{C}. \end{array}$$

The reader may note that, via the substitution $h_{pqi} = \lambda_{pqi}d_{pq}$ and (4.2), the problem [**RMP**] is equivalent to the arc-route formulation of the traffic assignment problem, with the exception that only a subset of the routes is available. The two formulations of the traffic assignment problem, either in terms of route flows, or in terms of link flows, differ only in the description of the polyhedron of feasible flows. From this simple observation, we may conclude that the constraints of the arc-route formulation in fact define an inner representation of the feasible set in the arc-node formulation of [**TAP**]. Moreover, the disaggregate simplicial decomposition algorithm is the link between simplicial decomposition and column generation methods for traffic assignment.

The DSD algorithm works as follows. Suppose that a disaggregate master problem, defined by a restricted set of routes, $\hat{\mathcal{R}}_{pq}$, $\forall (p,q) \in \mathcal{C}$, has been solved. Given this solution, the shortest-routes are calculated for all commodities. The sets $\hat{\mathcal{R}}_{pq}$ are augmented by the routes not contained in the sets already, and the procedure is repeated. The algorithm is valid; the proof given by von Hohenbalken (1977) may be applied, although his approach includes the dropping of all columns with zero weights.

The disaggregate master problem is a convex program with very simple linear constraints. In Larsson and Patriksson (1992), a scaled reduced gradient method is applied to [**RMP**]. The following description of the algorithm is taken from that reference.

Step 0 (*Initialization*) Let $k = 0$, and choose a point $\boldsymbol{\lambda}^0$ such that

$$\begin{array}{rcll} \displaystyle\sum_{i\in\hat{\mathcal{R}}_{pq}} \lambda^0_{pqi} & = & 1 & \forall (p,q)\in\mathcal{C} \\ \lambda^0_{pqi} & \geq & 0 \quad \forall i \in \hat{\mathcal{R}}_{pq} & \forall (p,q)\in\mathcal{C}. \end{array}$$

Step 1 (*Search direction generation*) Let

$$i_{pq} \in \arg\max_{i \in \hat{\mathcal{R}}_{pq}} \left\{ \lambda_{pqi}^k \right\}$$

be the basic variable index and let N_{pq} denote the set of nonbasic variables for commodity (p, q), respectively. Compute

$$\mathbf{r}_{pq} = \nabla_{N_{pq}} T\left(\boldsymbol{\lambda}^k\right) - \nabla_{i_{pq}} T\left(\boldsymbol{\lambda}^k\right) \cdot \mathbf{1},$$

where each component of $\nabla T\left(\boldsymbol{\lambda}^k\right)$ is calculated as

$$\frac{\partial}{\partial \lambda_{pqi}} T\left(\boldsymbol{\lambda}^k\right) = d_{pq} \sum_{a \in \mathcal{A}} \delta_{pqia} t_a \left(f_a^k\right).$$

(The sum in the right hand side is equal to the travel time on route i for origin-destination pair (p, q).) This calculation is made through the use of the relation (4.2), which defines the aggregate link flows f_a^k. The reduced gradient is the vector of all $\mathbf{r}_{pq}$. The search direction is defined as the negative of a scaled reduced gradient, that is,

$$\overline{r}_{pqi}^k = \begin{cases} -s_{pqi} r_{pqi} & \text{if} \quad i \neq i_{pq} \quad \text{and} \quad r_{pqi} \leq 0 \\ -\lambda_{pqi}^k s_{pqi} r_{pqi} & \text{if} \quad i \neq i_{pq} \quad \text{and} \quad r_{pqi} > 0 \end{cases}$$

$$\overline{r}_{pqi_{pq}}^k = -\sum_{i \neq i_{pq}} \overline{r}_{pqi}^k,$$

where $\mathbf{S} = \mathrm{diag}(s_{pqi})$ is a positive definite scaling matrix. (The use of the factors λ_{pqi}^k in the definition of $\overline{\mathbf{r}}_{pq}^k$ is a well known technique for deflecting the search direction away from near-active nonnegativity constraints. In our implementation, we have chosen the scaling coefficients $s_{pqi} = 1/d_{pq}$, $\forall i \in \hat{\mathcal{R}}_{pq}$. This choice of scaling coefficients corresponds to using a search direction $\overline{\mathbf{r}}$ based directly on the route travel costs.)

Step 2 (*Convergence test*) If $\overline{\mathbf{r}}^k = \mathbf{0}$, then terminate $\rightarrow \boldsymbol{\lambda}^k$ solves [**RMP**].

Step 3 (*Line search*) Let $\quad l_{\max} = \min_{(p,q) \in \mathcal{C}} \left\{ \min_{i \in \hat{\mathcal{R}}_{pq}} \left\{ -\frac{\lambda_{pqi}^k}{\overline{r}_{pqi}^k} \ : \ \overline{r}_{pqi} < 0 \right\} \right\}.$

Solve the line search problem

$$[\mathbf{LS}] \qquad \min_{l \in [0, l_{\max}]} \overline{T}(l) = T\left(\boldsymbol{\lambda}^k + l \cdot \overline{\mathbf{r}}^k\right),$$

which gives the solution l_k. (This line search is performed in total link flow variables by first calculating the link flows corresponding to $l_{\max}$.)

Step 4 (*New iteration point and convergence test*) Let $\boldsymbol{\lambda}^{k+1} = \boldsymbol{\lambda}^k + l_k \cdot \overline{\mathbf{r}}^k$.

Terminate if a stopping criterion is fulfilled $\rightarrow \boldsymbol{\lambda}^k$ solves [**RMP**] approximately.

Let $k := k + 1$.
Go to the search direction generation phase.

A main reason for using the disaggregate simplicial decomposition algorithm is its very good reoptimization capabilities. Reoptimization with respect to changes in link performance functions, travel demands and network topology are easily handled through appropriate modifications of the latest master problem, and its solution. The easy reoptimization is enabled due to the fact that more information is stored.

If the subproblem (3.10) is solved exactly, $L_c(\boldsymbol{\mu}^k)$ is a lower bound on the optimal objective value; see Section 3. Since, in general, the subproblems are only solved approximately, a lower bound on $L_c(\boldsymbol{\mu}^k)$ must be obtained. Because $L_c(\cdot, \boldsymbol{\mu}^k)$ is convex, the shortest-route problems solved repeatedly within the DSD algorithm define lower bounds on $L_c(\boldsymbol{\mu}^k)$, namely the well known Frank-Wolfe linearization bound. In each shortest-route subproblem we record the lower bound, and store the hitherto greatest of these. We denote this lower bound by LBD.

The lower bound is useful in a termination criterion only together with an upper bound on the optimal objective value. To this end, we consider adding to the augmented Lagrangean scheme a heuristic for generating feasible solutions from the approximate subproblem solutions.

4.3 Generating feasible solutions

The results of Theorems 3.1 and 3.2 are, obviously, not applicable if primal feasibility is desired finitely. In such cases, it is necessary to add to the augmented Lagrangean scheme some algorithm that heuristically converts a solution $\mathbf{f}^k$ into a primal feasible solution, $\overline{\mathbf{f}}^k$. There are many ways in which a feasible solution can be constructed from $\mathbf{f}^k$; important to note is that the structure of the feasible set must be exploited in the construction of the feasible solution, in order to make the heuristic computationally cheap, and that the heuristic should be conservative in the sense that $\mathbf{f}^k$ should not be altered too much in order to obtain primal feasibility.

To formalize the latter point, let $\overline{\mathbf{f}} = Pr(\mathbf{f})$ be the heuristic projection of $\mathbf{f} \in \mathbf{F}$ onto the feasible set $\mathbf{X} = \mathbf{F} \cap \{\mathbf{f} \mid \mathbf{f} \leq \mathbf{u}\}$. If the mapping Pr satisfies

$$\|Pr(\mathbf{f}) - \mathbf{f}\| \rightarrow 0 \text{ when } \mathbf{f} \rightarrow \mathbf{X}, \tag{4.3}$$

and if $\{\overline{\mathbf{f}}^k\} = \{Pr(\mathbf{f}^k)\}$ is a sequence generated by this algorithm within the augmented Lagrangean scheme, then

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

that is, under the assumption (4.3), $\{\overline{\mathbf{f}}^k\}$ is a sequence of feasible solutions, and $\{\overline{\mathbf{f}}^k\} \to \mathbf{f}^*$.

Let $\overline{\mathcal{A}}^k = \{a \in \mathcal{A} \mid f_a^k > u_a\}$ denote the set of arcs violating (4) in $\mathbf{f}^k$. The idea behind the heuristic, which is similar to that in Larsson and Liu (1989) for the linear multicommodity network flow problem, is to reduce the flows on these arcs, by iteratively transferring flows from a route in an origin-destination pair utilizing such arcs to other routes, within the same origin-destination pair, that are strictly feasible with respect to (4). This will obviously not destroy the feasibility with respect to the the other constraints; furthermore, this scheme is very simple to implement due to the disaggregated representation.

In the heuristic, origin-destination pairs are selected cyclically. Within a pair (p, q), a route $r \in \hat{\mathcal{R}}_{pq}$ is found, such that the route includes arcs in $\overline{\mathcal{A}}^k$, and a route $s \in \hat{\mathcal{R}}_{pq}$ is found, such that the route contains arcs that are all strictly feasible with respect to (4). A flow is then transferred from route r to route s so that either route r becomes feasible with respect to (4) or an arc contained in route r receives zero flow, or route s becomes saturated, i.e., the flow on one of the arcs contained in route s increases to its capacity. This is then repeated until no flow can be transferred using flows from commodity (p, q) only, and the next origin-destination pair is selected. The process terminates when the flows on all the arcs in $\overline{\mathcal{A}}^k$ have become feasible, or when no flow exchange is possible between any two routes in $\hat{\mathcal{R}}_{pq}$ for any origin-destination pair (p, q). (In the practical implementation, the feasibility procedure is terminated also if, in three consecutive main iterations, no infeasible arcs have become feasible.) In the latter case, the heuristic fails to obtain a feasible flow. Otherwise, a feasible flow, $\overline{\mathbf{f}}^k$, is obtained, and with it also an upper bound on the optimal value of [**TAP-C**]. We denote this upper bound by UBD.

From the construction of the flow exchange, the total infeasibilities decrease monotonically when sending flows along the cycles defined by the pair (r, s) of routes. Each of these cycles contains at most $|\mathcal{N}|$ arcs. Hence,

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

which implies that the total flow adjustment of the flow $\mathbf{f}^k$ made in order to obtain a feasible solution $\overline{\mathbf{f}}^k$ tends to zero when the initial total infeasibility tends to zero. Thus, the condition (4.3) is fulfilled, and we may conclude that the sequence $\{\overline{\mathbf{f}}^k\}$ is feasible, and optimizing in the limit (provided, of course, that the heuristic not ceases to find feasible solutions after some finite number of iterations).

4.4 Termination criteria

When the primal feasibility heuristic succeeds, both lower and upper bounds on the optimal value become available, and the most natural termination criterion is that the relative error,

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

is smaller than a prespecified parameter, $\varepsilon_1 > 0$.

In the early iterations of the augmented Lagrangean scheme, the feasibility heuristic may fail to generate any upper bound. The above termination criterion is then replaced by one defined by a measure of the infeasibility of $(\mathbf{f}^k, \boldsymbol{\mu}^k)$ with respect to the complementary slackness conditions

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

The termination criterion is that the maximal relative non-feasibility,

$$e_{max} = \max_{a \in \mathcal{A}} \left\{ \frac{|f_a^k - u_a|}{u_a} \;\middle|\; \mu_a^k > 0 \right\}, \tag{4.4}$$

is smaller than a prespecified parameter, $\varepsilon_2 > 0$.

4.5 Choice of penalty parameter values

The choice of the sequence $\{c_k\}$ of penalty parameters is obviously crucial to the practical efficiency of the algorithm, and some amount of experimentation must be made in order to yield a good range of their values. Guidelines are given for instance in Bertsekas (1982, Section 2.2.5). Due to the convexity of [**TAP-C**], the *threshold value*, $\underline{c}$, which c_k eventually must exceed, equals zero, which makes it possible to avoid ill-conditioning of the subproblems (3.10). On the other hand, the values must be large enough to yield a good dual convergence.

In the updating procedure suggested in Bertsekas (1982, p. 123), the penalty parameter is multiplied by a constant $\kappa > 1$ if the total infeasibility has not decreased by a factor $\gamma < 1$ between two successive iterations. That is,

$$c_{k+1} = \begin{cases} \kappa c_k & \text{if } \|\mathbf{g}^+(\mathbf{f}^k, \boldsymbol{\mu}^k, c_k)\| > \gamma \|\mathbf{g}^+(\mathbf{f}^{k-1}, \boldsymbol{\mu}^{k-1}, c_{k-1})\| \\ c_k & \text{otherwise,} \end{cases} \tag{4.5}$$

where $g_a^+(f_a, \mu_a, c) = \max\{g_a(f_a), -\mu_a/c\}$. Recommended values of the parameters κ and γ are $\gamma = 0.25$ and $2 \le \kappa \le 10$.

4.6 The augmented Lagrangean algorithm

We summarize below the steps of the algorithm and its convergence characteristics.

Step 0 (*Initialization*) Solve [**TAP**] approximately using the DSD algorithm. The solution is $\mathbf{f}^0$. Choose $\boldsymbol{\mu}^1$ according to (4.1), and $c_1 > 0$. Choose $\varepsilon_1, \varepsilon_2$. Let $k = 1$.

Step 1 (*Subproblem solution*) Solve the augmented Lagrangean subproblem (3.10) using the DSD algorithm, with $\mathbf{f}^{k-1}$ as starting point. The solution is $\mathbf{f}^k$. Update the lower bound LBD.

Step 2 (*Feasible solution generation*) Generate a feasible solution from $\mathbf{f}^k$ using the heuristic described in Section 4.3. The solution is $\overline{\mathbf{f}}^k$. Update the upper bound UBD.

Step 3 (*Convergence check*) If $\frac{UBD-LBD}{LBD} < \varepsilon_1 \rightarrow$ Stop. If upper bounds are not available, calculate e_{max} according to (4.4). If $e_{max} < \varepsilon_2 \rightarrow$ Stop. Otherwise, continue.

Step 4 (*Multiplier and penalty update*) Let $\boldsymbol{\mu}^{k+1}$ be given by (3.11) and c_{k+1} by (4.5). Let $k := k+1$, and go to Step 1.

The results of Bertsekas (1982, Section 2.2.5 and Corollary 5.9), Theorem 3.2, and the discussion made in Section 4.3 yield that the algorithm generates *bounded* sequences $\{c_k\}$, $\{\mathbf{f}^k\}$, $\{\overline{\mathbf{f}}^k\}$ (the latter assumed well defined), and $\{\boldsymbol{\mu}^k\}$ with the properties that

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

(2) $\{\overline{\mathbf{f}}^k\} \subset \mathbf{F} \cap \{\mathbf{f} \mid \mathbf{f} \leq \mathbf{u}\}$ and $\{\overline{\mathbf{f}}^k\} \rightarrow \mathbf{f}^*$,

(3) $\{\boldsymbol{\mu}^k\} \rightarrow \boldsymbol{\beta}$ (linearly) and $\{\mu_a^k\} \rightarrow 0$ (finitely), $\forall a \in \overline{\mathcal{A}}$.

5 Computational results

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 known uncapacitated test networks from the literature. In Table 5.1 we give the origins and sizes of the networks used.

City	Reference	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{C}\rvert$
	Nguyen and Dupuis (1984)	9	13	4
	Barton and Hearn (1979)	9	18	4
	Hearn and Ribera (1980)	9	18	4
Sioux Falls	LeBlanc *et al.* (1975)	24	76	528
Winnipeg	Nguyen (1976)	1052	2836	4344

Table 5.1: Test networks

In all these problems the travel time formulae may be written in the well known BPR form (Bureau of Public Roads, 1964)

$$t_a(f_a) = t_a^0 \left(1 + 0.15 \left(\frac{f_a}{c_a}\right)^{m_a}\right), \tag{5.1}$$

where t_a^0 is the free-flow travel time, $m_a \geq 1$, and c_a is the *practical capacity* of link a, that is, the flow at which the travel time starts to grow rapidly. The upper bounds u_a were in each problem instance constructed as

$$u_a = Kc_a, \; K > 0, \; \forall a \in \mathcal{A}, \tag{5.2}$$

i.e., as a uniform scaling of the practical capacities. For each of the test problems, the constant K, which we will sometimes refer to as the *capacity constant*, were chosen as small as possible, such that a smaller value of K would yield an infeasible problem; thus, difficult instances of [**TAP-C**] are created. (Note that, through the relations (1), (3) and (5.1), a uniform downward scaling of the vector $\mathbf{c}$ of practical capacities is equivalent to a uniform upward scaling of the demand vector $\mathbf{d}$. To choose the smallest possible value of K hence corresponds to uniformly increasing the demand for transportation to the level at which a further increase is not possible to assign onto the network due to the capacity constraints.)

In the next subsection we briefly discuss the parameters involved in the implementation of the proposed algorithm. Then, we compare its performance to that of pure Lagrangean relaxation and penalty methods for [**TAP-C**], when applied to an example network, and its efficiency when applied to the test problems given in Table 5.1. We also compare the augmented Lagrangean (3.3) used in our implementation to that used by Hearn and Ribera (1980). We finally investigate the characteristics of the algorithm and the optimal solution for varying values of the capacity constant K.

5.1 Implementational considerations

Some preliminary testing was made in order to determine overall best values of various parameters within the augmented Lagrangean scheme.

Since the DSD algorithm is executed several times during the course of the algorithm, the parameters given in Larsson and Patriksson (1992) needed to be modified somewhat. The acceptance parameter in the Armijo step length rule was, however, kept as in the DSD algorithm (i.e., 0.2 for the smaller problems and 0.3 for the larger ones), see Larsson and Patriksson. The same is true for the parameter that is used to terminate the solution of a disaggregate master problem. The relative accuracy demanded from each augmented Lagrangean subproblem (3.10) was, in most problems, 1.0%. This accuracy was sufficient to reach a final solution to within 0.1–1.0% of the optimal objective value, an accuracy which is considered sufficient for practical purposes. When demanding higher accuracies of the final solution, it was necessary to solve the subproblems more accurately. The strategy chosen in this case was to start with a relative accuracy of 1.0%, and then divide it by two for three main iterations.

The parameters, γ and κ, used in the updating formula (4.5) for the penalty parameter c_k, were chosen as $\gamma = 0.25$ and $\kappa = 5$, respectively. These values worked well on all test problems. The initial value of the penalty parameter, c_1, was chosen such that the sizes of the Lagrangean term and the penalty term in $\sum_a \overline{p}_a$ was roughly equal. With a much larger value of c_1, the penalty term would dominate, degrading the rate of convergence. A smaller value of c_1 turned out to also work well; very small values of the penalty parameters correspond to using Lagrangean relaxation, but with a slightly different updating formula for the multipliers than the traditional approach.

5.2 A comparison with the relaxation and penalty methods

The proposed method is a combination of Lagrangean relaxation and penalty approaches, which inherits the respective positive characteristics of these methods while avoiding their respective negative ones. To verify that the augmented Lagrangean scheme is superior, we compared the three methods' performance on the Sioux Falls network, demanding a high accuracy.

In Figure 5.1, the upper bounds for each method is given as a function of the computing time. The dashed line corresponds to the augmented Lagrangean method, the dashdotted line to the penalty function method and the dotted line to the pure Lagrangean relaxation method. Clearly, the augmented Lagrangean method converges faster to the optimum value (solid line) than the other two methods. The relative error is below 0.5% after less than 10 seconds. After 38.1 seconds of CPU time a relative error of 0.03% is obtained.

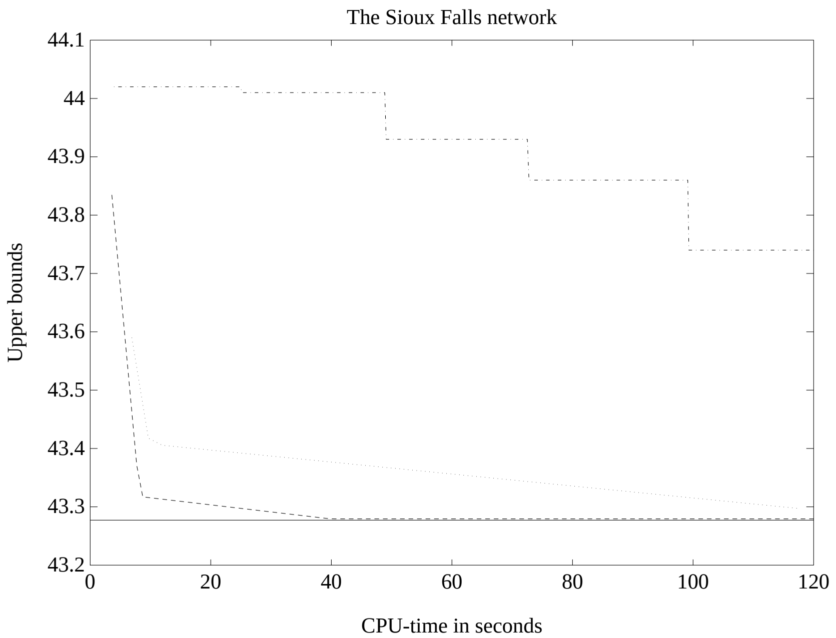

Figure 5.1: Comparing three relaxation methods

The penalty function method was coded using the quadratic penalty function (3.2) with $r_a \equiv 1$ and $m_a \equiv 2$, using the updating formula (4.5) for the penalty multiplier with the parameters $\kappa = 5$ and $\gamma = 0.25$. After 12.1 CPU seconds a solution with a relative error of 0.71% is found; reducing the relative error further to 0.17% takes 117 seconds.

The Lagrangean relaxation method was coded using the updating formula

$$\mu_a^{k+1} = [\mu_a^k + l_k g_a(f_a^k)]_+, \ \forall a \in \mathcal{A},$$

where l_k is the step size at iteration k. The step length rule is (Polyak, 1969)

$$l_k = \rho_k \frac{UBD - L_0(\boldsymbol{\mu}^k)}{\|\mathbf{g}(\mathbf{f}^k)\|_2^2}, \tag{5.3}$$

where L_0 is the Lagrangean dual function, $0 < \rho_k < 2$, and UBD is an upper bound on the optimal objective value. This step length rule was possible to utilize, since the feasibility heuristic was successful in each iteration. In practice, since the subproblem is not solved to optimality, $L_0(\boldsymbol{\mu}^k)$ is in (5.3) replaced by a lower bound, provided by the shortest-route solutions. In the experimentation, $\rho_k \equiv 1$ was set. After 120 CPU seconds and 791 iterations the relative error was 1.8%. The reason for the staircase structure of the upper bound is that the changes in the Lagrangean multipliers are in many iterations so small that the near-optimal solution to one subproblem is sufficiently close to optimality in the next subproblem to cause DSD to terminate immediately, thereby making no progress.

In the Lagrangean relaxation scheme, as well as in the augmented Lagrangean scheme, we recorded lower bounds on the optimal objective value. In Figure 5.2, lower and upper bounds for these two methods are depicted, with dashed and dotted lines corresponding to the augmented Lagrangean scheme and the Lagrangean relaxation scheme, respectively.

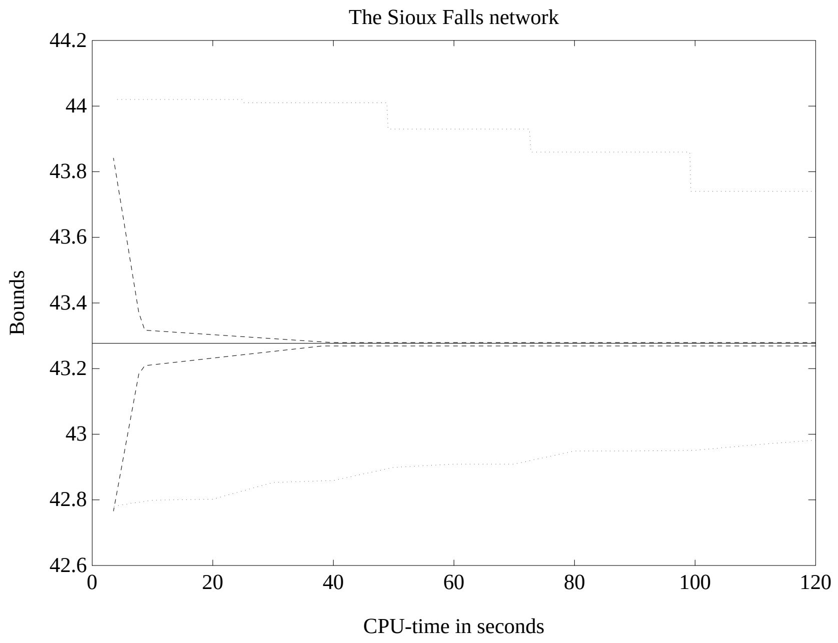

Figure 5.2: Comparing Lagrangean relaxation with the augmented Lagrangean method

From the figure, it is clear that for the Lagrangean relaxation scheme, the dual convergence is better than the primal one, although it is clearly inferior to that of the augmented Lagrangean scheme. Despite the fact that neither of the two methods are guaranteed to ascend the dual, in both cases the dual objective is monotonically increasing with iteration. In the case of the augmented Lagrangean scheme, this may be explained by the small values of the penalty parameters.

To conclude this section, we have thus verified that, not surprisingly, the augmented Lagrangean scheme outperforms the other two methods.

5.3 Further results for the augmented Lagrangean method

In this section, we present the results of numerical experiments with the augmented Lagrangean scheme performed on the networks given in Table 5.1. For each test network, we chose the smallest value of the capacity constant K in (5.2) possible. The number of link flows violating the capacity constraints in the solution of [**TAP**] is given for each problem. To determine the true deviation from the optimum given by the solution at termination, we solved each problem with very high accuracy, and recorded the optimal objective value. (The accuracy of the feasible solution obtained by the augmented Lagrangean scheme at termination is actually much better than is indicated by the relative accuracy given by the algorithm, due to the (relatively) poor quality of the linearization based lower bound.) For each problem instance, we report the computing time needed to obtain a given accuracy. To give an idea of the size of the master problems, we also present the number of routes generated and utilized by the algorithm.

The first network considered is a small network given in Nguyen and Dupuis (1984), with quadratic objective function ($m_a \equiv 1$). The constant $K = 5.5$ was chosen. When the uncapacitated problem is solved, three of the resulting link flows violate the capacity constraints. After four main iterations a feasible solution with an objective value of 85757.52 is obtained at termination, giving a relative error of 0.37%. The true deviation is, however, less than 0.0055%. The computing time used was 0.87 second and the average number of roads generated and utilized for each commodity was 2.75 and 2.25, respectively. Two of the three links that had non-feasible flows in the uncapacitated solution had flows on their upper bounds at the optimal solution for the capacitated problem.

The second network is a small quadratic network given in Barton and Hearn (1979). The constant $K = 1.5$ was chosen. Then, two of the link flows in the solution of the uncapacitated problem violate the capacities. After five main iterations and 0.60 CPU second a feasible solution with objective value 1483.22 and the relative error 0.49% was found. The true deviation from optimum was less than 0.041%, and the average number of generated and utilized routes were 6.75 and 5.75, respectively. In the optimal solution both of the initially non-feasible links had flows on their upper bounds.

The third network, given in Hearn and Ribera (1980), has the same basic characteristics as that given in Barton and Hearn (1979), apart from the travel time formula; for instance, $m_a \equiv 4$. In Hearn and Ribera, the constant $K = 1$ is considered. To solve this problem, Hearn and Ribera apply an augmented Lagrangean scheme, with an augmented Lagrangean given by Pierre and Lowe (1975). There,

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

whereas (3.4) is used in our algorithm. The augmented Lagrangean of Pierre and Lowe was tested in our implementation; the result showed that the performance of the two methods was very similar in this application. Another, more important, difference is that Hearn and Ribera apply the Frank-Wolfe method on the subproblems. The method described in this paper gave after 25 main iterations and 2.9 seconds of CPU time an objective function value of 2307.91. The feasibility heuristic was not in any iteration able to find a feasible solution, and the relative accuracy is therefore never calculated. The termination criterion was instead that the maximum relative non-feasibility (4.4) was small. The results from the two algorithms are compared in Table 5.2, where, in columns A and B, respectively, we present the link flows produced by the algorithm of Hearn and Ribera, and the method (3.10), (3.11).

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 5.2: Comparing two augmented Lagrangean algorithms

The fourth network is a model of the city of Sioux Falls, South Dakota, and is given in LeBlanc *et al.* (1975). The capacity constant $K = 2.0$ is used; the link performance functions are of the form (5.1), with $m_a \equiv 4$. After two main iterations the algorithm terminates with an objective value of 43.371. The relative error was 0.43% while the true deviation from optimum is less that 0.22%. The CPU time used was 7.6 seconds and the number of routes generated and utilized was 2.23 and 1.47 on the average for each commodity. When solving the uncapacitated problem, 14 links had flows which were non-feasible, and all of these links had flows on their upper bounds at termination. In Figure 5.3 the upper (solid line) and lower bounds (dashed line) are given.

The fifth network, given in Nguyen (1976), models the city of Winnipeg. The constant $K = 1.8$ is used; the travel time formulae have varying values of m_a. After 776 CPU seconds and 12 main iterations an objective value of 892201.2 was found. The relative error was 0.72%. Only in the last main iteration did the feasibility heuristic succeed in obtaining a feasible flow. On the average 5.19 routes were generated in an origin-destination pair and 3.34 of these were utilized. In the uncapacitated solution, 12 link

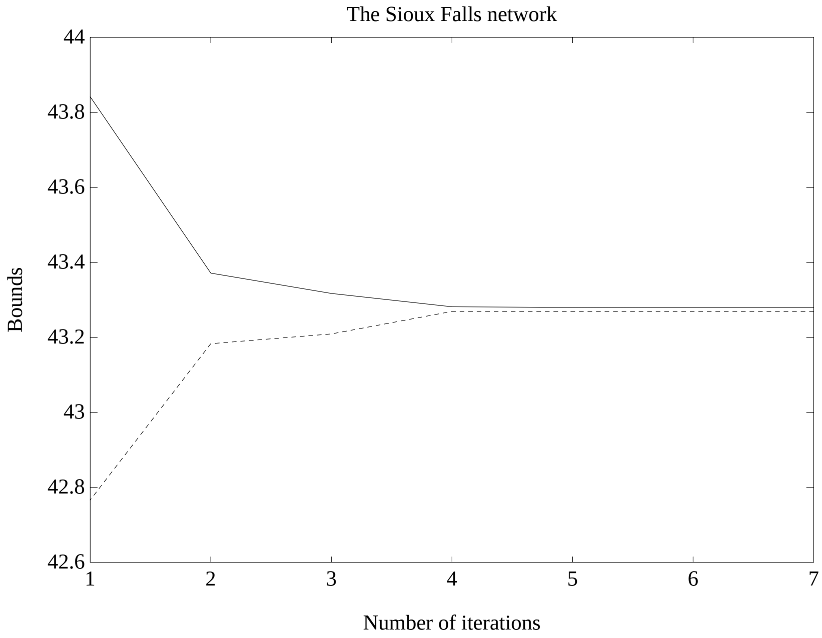

Figure 5.3: Upper and lower bounds versus iterations

flows violated the capacity constraints; all of these had flows at their upper bounds in the final solution.

The experiments presented above indicate that capacitated traffic assignment problems can be effectively solved by an augmented Lagrangean scheme. To reach a near-optimal solution, only a factor of three to four of the solution time required for solving the uncapacitated model in the initialization step of the algorithm is needed. This is in part due to the availability of a good starting solution for the Lagrange multipliers, but also to the good reoptimization facilities and utilization of the problem structure of the DSD algorithm.

To further investigate the behaviour of the algorithm for problems with varying characteristics, we solved the test network of Hearn and Ribera for various values of the capacity constant K, i.e., we varied the capacities through a uniform scaling. By inspection it is possible to see that a constant $K < 1$ yields an infeasible problem. For each value of $K \geq 1$ tested, we recorded at termination the objective value, the computing time used, the number of routes generated and used and the travel times on the utilized routes.

We present in Figure 5.4 the objective value versus the capacity constant; we note that for $K \geq 3.4$ the problem is practically uncapacitated.

In Figure 5.5 we present the computing times used for various values of the capacity constant. The non-monotonicity is due to the fact that the problems are not solved exactly. In this case the termination criterion was that the maximum relative non-feasibility is low.

In Figure 5.6 we present the average travel time (solid line) and generalized travel time (dashed line) for a utilized route in the solution of [**TAP-C**], as a function of K. Although the average travel time can be expected to decrease when the network is improved by an

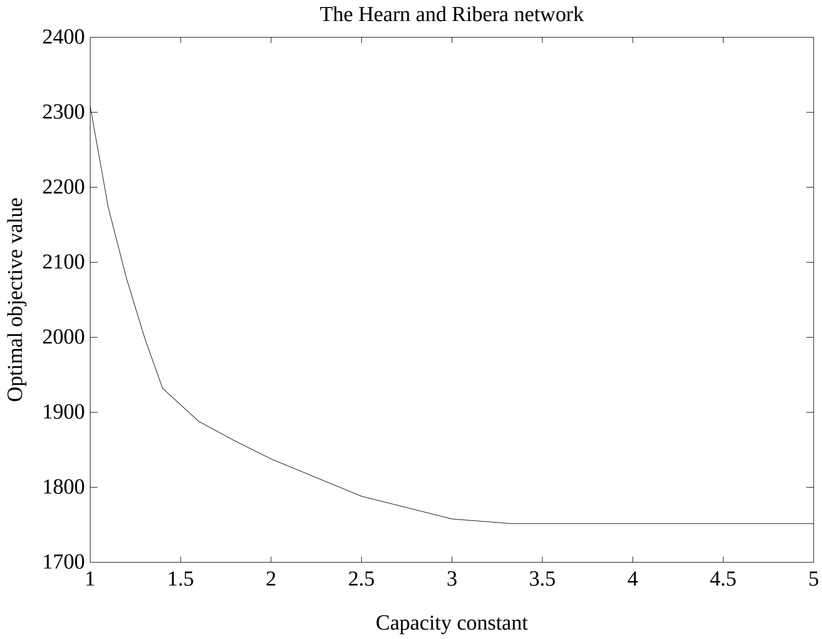

Figure 5.4: Optimal objective function value versus the capacity constant

increased capacity, this is not necessarily the case, as the figure shows. This phenomenon is known as *Braess' paradox*; see, e.g., Sheffi (1985, pp. 75–77).

We finally present the average number of routes generated and utilized for different values of the capacity constant. Since the number of routes generated equals the number of variables in the master problems, it gives an idea of the average size and computational difficulty of the subproblems. In Figure 5.7, the solid line corresponds to utilized routes and the dashed line to the generated routes. The number of routes generated and utilized increase rapidly for values of the capacity constant below 1.6, indicating a very high congestion.

6 Conclusions

The major contribution of this work is that we show that the traffic assignment model may remain computationally tractable even when it is extended through the addition of side constraints. We also make some additional, technical, contributions. First, under a weak regularity assumption, we establish a linear rate of convergence for the sequence of multiplier iterates. Second, we propose an advanced choice of initial multiplier values, and, third, we present a heuristic method for finding feasible flow patterns by carefully manipulating the augmented Lagrangean subproblem solutions, which are, in general, infeasible in the original problem.

Summarizing the results, the proposed augmented Lagrangean solution method was found to be an effective and efficient computational means for handling the capacity constraints. A comparison with the computing times for the uncapacitated problem in the initialization

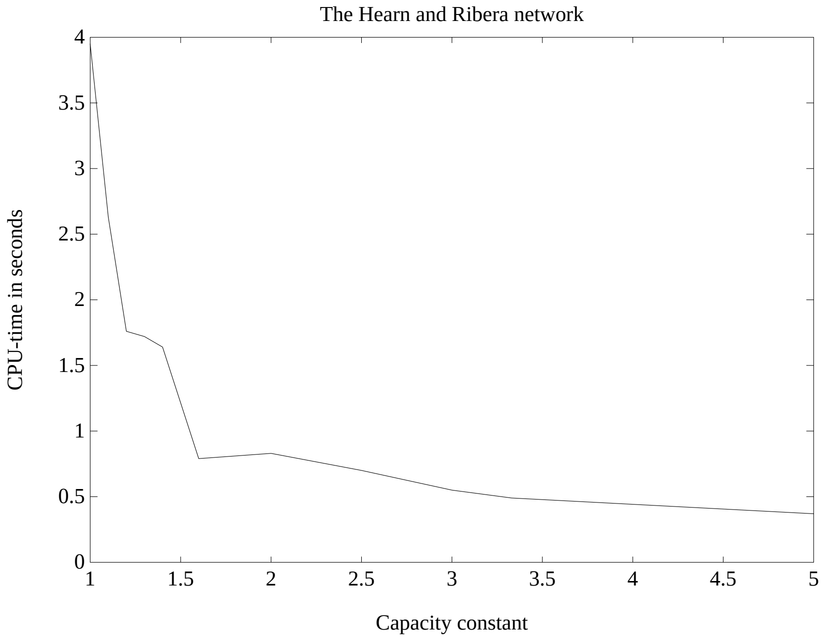

Figure 5.5: Computing times versus the capacity constant

of the algorithm yields that the capacitated problem takes only around three to four times longer time to solve. We consider this to be a quite modest increase in computing time since the very profitable Cartesian product structure of the feasible set of [**TAP**] is lost through the introduction of coupling capacity constraints, thus making [**TAP-C**] a significantly more difficult problem.

One explanation for the good performance of the proposed method is the fast convergence of the multiplier iteration of the augmented Lagrangean scheme; as compared to the convergence of the subgradient method applied to the ordinary Lagrangean dual, the difference is considerable. We have also observed that the fast convergence of the multiplier iteration is gained also when the subproblem is only mildly penalized, i.e. when the penalty term of the augmented Lagrangean function is small compared with its Lagrangean term. As a rule of thumb, we suggest choosing the initial value of the penalty parameter so that the penalty and the Lagrangean terms, respectively, are of about the same magnitude. Furthermore, the choice of initial values for the Lagrangean multipliers according to (4.1) was found to be quite satisfactory, since the links with capacity restrictions that were violated in the resulting subproblem solution in most cases had small excess flows.

The efficiency of the disaggregate simplicial decomposition method for solving the uncapacitated traffic assignment problem (see Larsson and Patriksson, 1992) motivated its use for solving the subproblems. Furthermore, because of its very good reoptimization capabilities it is very suitable for solving a sequence of subproblems in an iterative scheme, and the computational effort needed for solving the subproblems was indeed observed to be reduced for every main iteration of the augmented Lagrangean method. Different strategies for terminating the solution of the subproblems were tested. Clearly, there is a trade-off between the accuracy to which the subproblems are solved and the number of

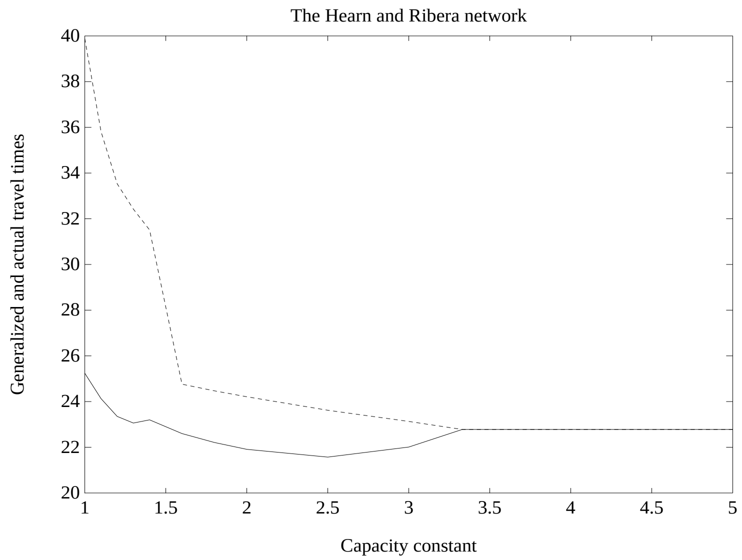

Figure 5.6: Travel times versus the capacity constant

main iterations that are needed to reach convergence in the overall scheme; our experiments indicate that it is, from a practical point of view, not worthwhile to demand a high accuracy when solving the subproblems, at least not in the early iterations, even though the low accuracy results in more main iterations.

Regarding the quality of the global upper and lower bounds on the optimal value of [**TAP-C**], it was found that, whenever the heuristic procedure succeeded in finding a feasible solution, the upper bound was of a higher quality than the corresponding lower bound. (In Larsson and Patriksson, 1992, it was established that the quality of the lower bound on the optimal value of an uncapacitated traffic assignment problem obtained from a shortest-route subproblem is, in general, rather poor, and this property is transferred to the subproblem of the augmented Lagrangean method.) However, the lower bound of the augmented Lagrangean method is much stronger than the one provided by the ordinary Lagrangean relaxation (see Figure 5.2).

The results of the experiment with differently strong link capacities were the expected ones; a tightly capacitated problem requires a longer solution time since more routes are generated and utilized. We also observed the so-called Braess' paradox, i.e., that tighter capacities may allow shorter average travel times.

There are many interesting and promising issues for further research along the lines initiated by this work. First of all, the augmented Lagrangean scheme can be varied, and certainly also enhanced, in several ways. One may for instance employ more elaborate updating formulae for the penalty parameter (see e.g. Bertsekas, 1982), or penalize constraints using different, and individually updated, parameters. Another possibility is to scale the constraints, e.g. 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 augmented Lagrangean dual function through the transformation of the multiplier space with a diagonal matrix; when the multiplier up-

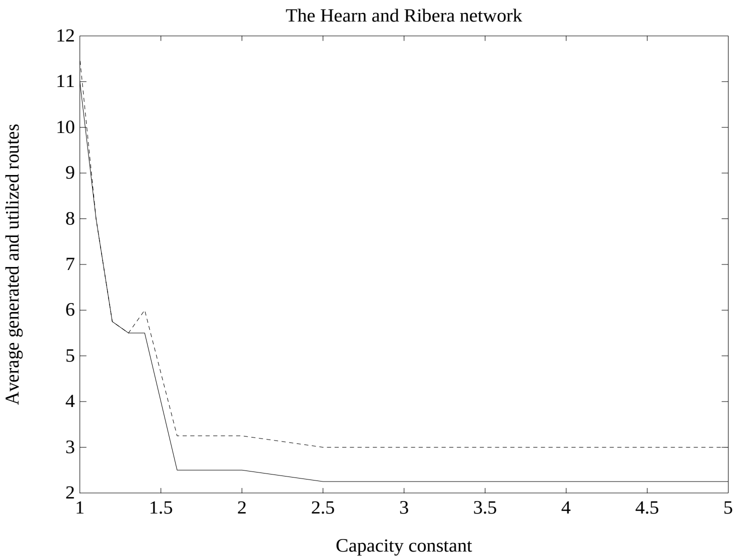

Figure 5.7: Average generated and utilized routes versus the capacity constant

dating formula (3.11) is applied in the transformed multiplier space, the resulting step direction will therefore in the original multiplier space correspond to a direction obtained through dilating the one defined by (3.11) with a diagonal matrix. Still other opportunities for improvements are the application of more advanced multiplier iteration formulae (see e.g. Bertsekas, 1982), e.g. approximate second-order formulae, or other types of augmented Lagrangean functions (e.g., Tseng and Bertsekas, 1993) than the ones used in our development.

Further, the heuristic procedure for generating feasible solutions to [**TAP-C**] can be designed in many different ways, and there may very well be strategies that work better than the one used in our current implementation; we believe that one may probably design a more efficient heuristic procedure by manipulating the infeasible link flow solution directly, instead of the route flow solution, and also taking the link costs into account when striving for feasibility.

As a subject for further research it would be interesting to investigate the application of the proposed 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, possible 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 may conjecture that such side constrained traffic assignment models can be efficiently dealt with computationally, although the Cartesian product structure is destroyed by the side constraints, but it is of course necessary to evaluate the proposed solution strategy when applied to models with various kinds of more complex side constraints in order to establish this conjecture as a fact. (The reader should note that the strategy of extending the basic model with side constraints gives a large flexibility in the construction of the model, since the side constraints are allowed to be non-linear or non-separable.) Such a result would hopefully encourage the research community to a renewed interest into the art of modelling real-world traffic problems.

As described at the end of Section 2.2, explicit link capacities may be used by the traffic engineer as a means for calculating proper 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 interactions and restrictions into account. In such a procedure one would formulate and solve a traffic assignment problem including a set of suitable side constrains, and then utilize the Lagrangean multipliers for these constraints to derive adjusted travel time functions which indirectly describe the additional model components. This way of deriving complex travel time functions may be to prefer to a calibration of parameters in the travel cost functions, since it may be comparably easy to identify a set of appropriate side constraints and estimate proper values of their coefficients, which may have very tangible physical interpretations.

Acknowledgements

The research leading to this report was financially supported by the Swedish Research Council for Engineering Sciences (TFR) and The Swedish Transport and Communications Research Board (KFB). M.Sc. Håkan Fortell implemented the proposed solution method and made the computational experiments as a part of his final project.

References

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

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

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

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

[5] Bazaraa, M.S. and Shetty, C.M., *Nonlinear Programming: Theory and Algorithms.* John Wiley & Sons, New York, NY, 1979.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[21] Hearn, D.W. and Lawphongpanich, S., *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, 1989.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[36] Lasdon, L.S., *Optimization Theory for Large Systems.* MacMillan, New York, NY, 1970.

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

[38] Luenberger, D.G., *Linear and Nonlinear Programming.* Addison-Wesley, Reading, MA, second ed., 1984.

[39] Miller, S.D., Payne, H.J. and Thompson, W.A., *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.

[40] Nagurney, A., *Network Economics: A Variational Inequality Approach.* Kluwer Academic Publishers, Dordrecht, 1993.

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

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

[43] Patriksson, M., *The traffic assignment problem—theory and algorithms.* Report LiTH-MAT-R-90-29, Linköping Institute of Technology, Linköping, Sweden, 1990.

[44] Payne, H.J. and Thompson, W.A., *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.

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

[46] Pigou, A.C., *The Economics of Welfare.* MacMillan & Co., London, 1920.

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

[48] Polyak, B.T., *Minimization of unsmooth functionals.* USSR Computational Mathematics and Mathematical Physics **9**, pp. 14–29, 1969.

[49] Polyak, B.T., *Introduction to Optimization.* Optimization Software, New York, NY, 1987.

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

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

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

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

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

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

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

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

[58] Smith, M.J., *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, 1987.

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

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

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

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

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