Nonlinear programming advances in mathematical programming with complementarity constraints

By Daniel Ralph
28 September 2007
Judge Business School, University of Cambridge, Cambridge CB2 1AG, UK

Given a suitably parameterised family of equilibrium models and a higher level criterion by which to measure an equilibrium state, mathematical programs with equilibrium constraints (MPECs) provide a framework for selecting good or even optimal parameters. An example is toll design in traffic networks, which attempts to reduce total travel time by choosing which arcs to toll and what toll levels to impose. Here, a Wardrop equilibrium describes the traffic response to each toll design. Communication networks also have a deep literature on equilibrium flows that suggest some MPECs. We focus on mathematical programs with complementarity constraints (MPCCs), a subclass of MPECs for which the lower level equilibrium system can be formulated as a complementarity problem. An MPCC can be immediately written as a nonlinear program (NLP) but, regrettably, the constraints of the latter lack some stability properties that are considered essential in analysing and solving NLPs. A related issue is that MPECs and MPCCs are generally nonconvex, which is a consequence of the upper level objective clashing with users' objectives in the lower level equilibrium program. While global optimisation of MPECs and MPCCs is in a state of infancy, the last decade of research has paved the way for finding local solutions of MPCCs via standard NLP techniques if not off-the-shelf NLP software. We give a selective review of these advances. Special attention is paid to Lagrangian techniques which have become essential in describing what "stationary" means for MPCCs, and how to solve MPCCs computationally.

Keywords: MPECs, MPCCs, complementarity problems, equilibrium constraints, variational inequalities, inverse optimisation, system design, traffic networks, communication networks

1. Introduction

Mathematical programs with equilibrium constraints (MPECs) are a class of optimisation problems of the form

$$\begin{array}{ll} \min\limits_{x,y} & f(x,y) \\ \text{subject to} & (x,y) \in Z \;=\; \text{standard region, e.g., polyhedron} \\ & y \text{ solves an equilibrium problem that depends on } x \end{array} \tag{1.1}$$

where $x \in \mathbb{R}^n$ may variously be called the upper level or leader or design or control vector, and $y \in \mathbb{R}^m$ the lower level or follower or response or state vector, respec-

tively; the upper level objective function $f : R^{n+m} \to R$ is smooth; and lower-level optimization or equilibrium problem is formulated with smooth functions.

Equilibrium constraints may come in the form of a game, a variational inequality (VI) or stationary conditions of one of these problems or of an optimisation problem. In this paper we restrict our attention to equilibrium constraints that can be formulated via finitely many smooth equations, inequalities and complementarity relations, namely, in the framework of an underdetermined mixed complementarity problem (MCP). That is, the MPEC will be a mathematical program with complementarity constraints (MPCC) as follows:

$$\begin{array}{lll} \min\limits_{z} & f(z) & \\ \text{subject to} & g(z) \geq 0, & h(z) = 0 \\ & 0 \leq r(z) \quad \perp & s(z) \geq 0 \end{array} \tag{1.2}$$

where $r, s : \mathbb{R}^n \to \mathbb{R}^m$, $g : R^n \to \mathbb{R}^{m_g}$, $h : R^n \to \mathbb{R}^{m_h}$ are smooth and $\perp$ denotes orthogonality. Here we do not distinguish between upper and lower level variables, rather aggregate them into the single vector z. The MPCC format is important for accessing many nonlinear programming techniques.

Note that stationary conditions of nonlinear programs (NLPs) are formulated as Karush-Kuhn-Tucker, KKT, systems (Nocedal and Wright, 1999), which are themselves mixed complementarity problems. For example we may reformulate a bilevel optimisation problem as an MPCC by replacing the lower level problem with its KKT conditions. This is attractive particularly when the lower level problem is convex, and its stationary conditions characterise global optimality.

General references on MPECs include the monographs Luo *et al.* (1996) and Outrata *et al.* (1998), and the more recent bibliography Dempe (2003) on the bilevel programming and MPEC literature.

This paper highlights advances in the application of nonlinear programming techniques to MPCCs against the background of equilibrium problems in traffic and communication networks. Special attention is paid to Lagrangian or KKT conditions and related algorithms, which have become increasingly useful in providing robust and large scale solution methods for finding local solutions of MPCCs.

A cautionary note is that MPCCs are nonconvex optimisation problems, and the computational techniques that reviewed here are aimed at finding stationary points of MPCCs rather than global optima. Nonconvexity appears in the simplest complementarity constraint $0 \leq x \perp y \geq 0$, over scalars x and y. This gives the nonconvex feasible set $\{(x, 0) : x \geq 0\} \cup \{(0, y) : y \geq 0\}$ which is the union of two polyhedral convex "pieces". There is a dearth of research on global optimisation of MPECs though we can point to Hu *et al.* (2007) for exact methods and Sumalee (2004a,b) for heuristic search methods, and references therein.

Section 2 reviews some basic equilibrium and MPEC/MPCC models in traffic and communication networks. The next two sections consider NLP approaches to generic MPCCs. Section 3 shows how the MPCC-Lagrangian, which is adapted directly from classical nonlinear programming, provides stationary conditions for MPCCs. The so-called strong stationary conditions are especially important because they are often "tight" and can be checked easily, circumventing the combinatorial nature of MPCCs. Section 4 presents several NLP approaches to solving MPCCs, each of which can be analysed in terms of the MPCC-Lagrangian. The

fact that superlinear convergence to local solutions can be achieved by two of these methods supports the claim that NLP algorithms hold great promise in tackling MPCCs. Section 5 points to future work by discussing some related literature: the price of anarchy and equilibrium programs with equilibrium constraints.

2. From equilibrium models to MPECs to MPCCs

Consider an equilibrium system, or stationary conditions of an optimisation problem, that we call the "forward model". The data of such a model implicitly describe a "forward output", namely an equilibrium solution. There are two paradigms fundamental to scientific modelling that give rise to MPECs when the forward model is an equilibrium problem.

The first paradigm is the design problem, in which parameters that effect the equilibrium model are adjusted to explore the space of achievable equilibria, and thereby improve the performance of an existing or planned system. The second paradigm is the inverse process of parameter estimation, where we try to identify an existing equilibrium system with one that appears in a parameterised family. An inverse MPEC could be formed by minimising the least squares distance between the given observations of the existing equilibrium state and the corresponding values of the forward equilibrium, subject to changing the model parameters within acceptable limits. This can also be regarded as calibration.

The goal of this section is to briefly describe some important forward equilibrium models in road traffic networks, in section (a), and communication networks, in section (b), and discuss associated design and inverse MPECs. It will be important for us to show the transformation of an equilibrium model into an MCP, so that subsequent MPECs can be written as an MPCCs. Section (c) concludes with a brief discussion of different directions taken in the research literatures on equilibria in traffic networks and communication networks.

(*a*) *MPCCs based on traffic equilibria*

(i) *Forward model of Wardrop traffic equilibria*

Notation: Consider a network with N nodes, a set of directed arcs or links $\mathcal{A}$, where links are denoted either ij for flow from node i to j or more generally a, a set $\mathcal{O}$ of origin-destination (OD) pairs, denoted kl or κ, and corresponding list of OD demands or quantities $q_\kappa \geq 0$, $\kappa \in \mathcal{O}$. Each OD pair kl defines a commodity in a multi-commodity network flow problem, whose feasible set is given by those nonnegative vectors $x^{kl} = (x_a^{kl})_{a\in\mathcal{A}}$ that achieve a transfer of q_{kl} units from k to l. Elastic demand, that decreases as congestion increases, can easily be adapted into this approach. See the monographs of Patriksson (1994) and Sheffi (1985) for general views of traffic modelling.

Wardrop traffic equilibria: We wish to find the equilibrium pattern of traffic flow $f = \sum_{\kappa\in\mathcal{O}} x^\kappa$ over feasible OD flows x^κ, bearing in mind that travel durations on links depend on link flows. Additional constraints, such as as upper bounds on any link flow f_a, can also be imposed to define the set of feasible flows f. Wardrop's equilibrium principle (Wardrop, 1952) states that, at equilibrium, OD traffic will only flow on shortest time paths. The shortest path idea generalises directly to a

cost-of-travel function that may combine any number of factors such as distance, time, fuel cost, etc.

VI formulation of traffic equilibrium: Smith (1979), followed by Dafermos (1980), characterises Wardrop's traffic equilibrium as a variational inequality, as follows. Let $d_a(f)$ denote the unit or marginal travel time, or duration for short, of flow on link a given a feasible flow vector f.† An equilibrium is a feasible flow f^* such that for any other feasible flow f,

$$\sum_{a\in\mathcal{A}} d_a(f^*)f^*_a \;\le\; \sum_{a\in\mathcal{A}} d_a(f^*)f_a. \tag{2.1}$$

We denote this problem by VI(d, F) where F is the set of feasible flows and d is the vector over links of travel durations, $d(f) = (d_a(f))_{a\in\mathcal{A}}$. Formally, VI$(d, F)$ is the problem of finding $f^* \in F$ such that $d(f^*)^\top(f - f^*) \ge 0$ for all $f \in F$. A computational attraction of the VI formulation over the definition of Wardrop equilibrium is that the former is based on link flows, while the latter requires enumeration of paths, which can grow least exponentially in the number of links.

MCP formulation of traffic equilibrium: We express F as a polyhedral set via commodity link flows, as

$$\left.\begin{array}{rcl} f & = & \displaystyle\sum_{\kappa\in\mathcal{O}} x^\kappa \\ x^\kappa & \ge & 0 \\ Ax^\kappa & = & b^\kappa \end{array}\right\} \;\forall \kappa \in \mathcal{O} \tag{2.2}$$

where $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$, $A \in \mathbb{R}^{N\times|\mathcal{A}|}$ is the node-arc incidence matrix, whose entry in row k and column $\kappa = ij$ is 1 if $j = k$, -1 if $i = k$ and 0 otherwise; and $b^{kl} \in \mathbb{R}^N$ consists of net inflows at each node, i.e., its ith component is $-q_{kl}$ if $i = k$, q_{kl} if $i = l$ and zero otherwise. We introduce a dual multiplier $\mu = (\mu^\kappa)_{\kappa\in\mathcal{O}}$ where each $\mu^\kappa \in \mathbb{R}^N$ corresponds to the OD flow balance equation $Ax^\kappa = b^\kappa$. The dual condition that characterises f^* as a solution of VI(d, F) is the existence of a multicommodity flow vector x^* such that (2.2) holds at $(f, x) = (f^*, x^*)$, and KKT multiplier μ such that‡

$$0 \le d\,(f^*) + A^\top\mu^\kappa \quad \perp \quad x^\kappa \ge 0, \quad \forall \kappa \in \mathcal{O}. \tag{2.3}$$

That is, (2.3)-(2.2) is an equivalent MCP formulation of a Wardrop equilibrium.

This kind of equivalence between a VI and its corresponding MCP relies on, first, the feasible set F of the former being written in terms of finitely many equalities and inequalities, second, closedness and convexity of F, and third, a condition on F at a prospective equilibrium f^* called a constraint qualification, that allows the geometric tangent cone to F at f^* to be described by linearising the constraints that are "active" there. For polyhedral convex sets as discussed above, this constraint qualification holds automatically. See Facchinei and Pang (2003) for details.

† Often each d_a only depends on flow on arc a, which is convenient for optimisation reformulations and analysis as discussed later.

‡ More traditionally, nonnegative multipliers associated with the inequalities $x^\kappa \ge 0$ would also appear in the complementarity formulation. These, however, can be eliminated in the complementarity conditions to give the equivalent dual conditions (2.3).

Under strict monotonicity¶ or other conditions on d (Facchinei and Pang, 2003), the solution f^* is unique though x^* or μ may not be. It turns out for a connected graph that the matrix A has rank $N-1$, and that elimination of any row from each of the constraints $Ax^\kappa = b^\kappa$, $\kappa \in \mathcal{O}$, preserves the feasible set. It follows that a KKT point (x^*, μ) is associated with a unique KKT multiplier $\bar{\mu}$ such that $\bar{\mu}^\kappa$ is independent of κ and $\bar{\mu}^\kappa_N = 0$.

A final comment is that the number of variables in the MCP formulation is $|\mathcal{O}|(|\mathcal{A}| + N)$, on the order of $N^2|\mathcal{A}| + N^3$ if there is OD demand from every node to every other node. This is considerably larger than the number of flow variables needed for the VI formulation but still much less in general than the possible number of routes needed for the definition of a Wardrop equilibrium.

Existence of traffic equilibria: Provided F is convex, compact and nonempty, and d is continuous on F, it follows from Brouwer's fixed point theorem that $\mathrm{VI}(d, F)$ has a solution (Facchinei and Pang, 2003), hence that the associated KKT condition also has a solution. It is clear that F is closed and convex. Note that F is bounded since for each $kl \in \mathcal{O}$, the flow on any path from k to l is nonnegative and bounded above by q_κ; thus each link flow f_a lies between 0 and $\sum_{\kappa \in \mathcal{O}} q^\kappa$. F is also nonempty if there are no additional constraints beyond feasibility of the subordinate OD flows, and the network is connected. These properties and hence existence of a traffic equilibrium are preserved by many additional constraints such as sufficiently large upper bounds u_a on total flow on link a. The first existence and uniqueness proof of Wardrop equilibria is via an entirely different route however, namely the equivalent optimisation reformulation of Beckmann *et al.* (1956). This equivalence relies on a symmetry condition on the link cost functions that holds trivially if the cost of flow on link a depends only on flow on that link.

(ii) *Toll design in traffic networks*

Think of decreasing total travel time in road traffic network by choosing an appropriate cordon charging scheme around a city, of which London's cordon charging scheme is an example. The currency and challenges of road tolling in practice are well described in Kelly (2006).

Some motivation for tolling schemes comes from the observation that the maximum welfare attainable under a centralised planner (benevolent dictator), who prescribes OD routes to all road users, can be achieved by a traffic equilibrium under a tolling scheme where there is a possibly different marginal toll cost on every link. This result assumes that all users view the tradeoff between time and toll cost in the same way, i.e., share the same time-value of money, an assumption that will be preserved below in order to conserve notation. Accommodating user classes each with a distinct time-value of money is possible by introducing further commodities into the network flow model.

Of course tolling every link would be difficult, moreso for link-dependent toll levels. Cordons reduce the complexity of the toll scheme by setting a single toll price throughout the cordon and a zero toll elsewhere. A cordon is essentially a connected subgraph of the network. For simplicity we will not model this here — see Sumalee (2004a,b) for a graph theoretic description that also converts to a computational

¶ The vector function $d(f)$ is monotone on F if $(d(f) - d(f'))\top(f - f') \geq 0$ for any $f, f' \in F$, and strictly monotone if the inequality holds strictly when, in addition, $f \neq f'$.

description — but consider the related problem (Sumalee, 2004a) of choosing a subset $\hat{\mathcal{A}}$ of all links and a single toll charge τ applied to each unit of traffic on those links, which constitute the network design variables, in order to maximise a system utility function.

Suppose travel along link a is tolled at £τ per unit flow. We model the user cost function of using link a as $\gamma d_a(f) + \tau$, where $d_a(f)$ is link duration, as above, and $\gamma > 0$ scales time to money. Given the arc flows f, list of toll arcs $\hat{\mathcal{A}}$ and toll level τ, define

$$c_a(f, \hat{\mathcal{A}}, \tau) = \begin{cases} \gamma d_a(f) + \tau, & \text{if } a \in \hat{\mathcal{A}}, \\ \gamma d_a(f), & \text{otherwise,} \end{cases}$$

and $c(f, \hat{\mathcal{A}}, \tau) = (c_a(f, \hat{\mathcal{A}}, \tau))_{a \in \mathcal{A}}$. The system utility associated with any flow f is the total benefit, $\tau \sum_{a \in \hat{\mathcal{A}}} f_a$, less total cost†, $\sum_{a \in \mathcal{A}} c_a(f, \hat{\mathcal{A}}, \tau) = \gamma \sum_{a \in \mathcal{A}} d_a(f) + \tau \sum_{a \in \hat{\mathcal{A}}} f_a$. Since toll revenues and toll costs net to zero, maximising system utility is equivalent to minimizing total user travel time,

$$\begin{array}{ll} \min\limits_{\hat{\mathcal{A}}, \tau, f} & \sum\limits_{a \in \mathcal{A}} d_a(f) f_a \\ \text{subject to} & \hat{\mathcal{A}} \subset \mathcal{A} \\ & \tau \geq 0 \\ & f \text{ solves VI}(c(\cdot, \hat{\mathcal{A}}, \tau), F). \end{array}$$

The combinatorial constraint $\hat{\mathcal{A}} \subset \mathcal{A}$ makes this problem even more difficult than a standard MPEC whose variables are continuous. Nevertheless genetic algorithms are an effective heuristic for finding good toll designs (Sumalee, 2004a,b).

To pose the toll design problem as an MPCC, observe that the MCP corresponding to the equilibrium constraint is (2.2) and (2.3) with ∇d_a replaced by $\nabla c_a(\cdot, \hat{\mathcal{A}}, \tau)$ in the latter.

(iii) *OD demand estimation, an inverse traffic equilibrium problem*

Consider the forward traffic equilibrium model given above where the origin-destination demand vector $q = (q_\kappa)_{\kappa \in \mathcal{O}}$ is unknown, but we believe it lies in a closed convex set Q that is based on values for OD demands that are known from a previous period. Suppose further that we have measured arc flows f^*_a in the current period, for a in a subset $\bar{\mathcal{A}}$ of $\mathcal{A}$, where traffic flow is assumed to be at equilibrium. We now wish to calibrate the network model by estimating OD demands q for the current period. This is a well studied problem, see Chen and Florian (1998) and references therein.

A least squares version of the calibration problem is

$$\begin{array}{ll} \min\limits_{d, f} & \sum\limits_{a \in \bar{\mathcal{A}}} (f_a - f^*_a)^2 \\ \text{subject to} & q \in Q \\ & f \text{ is an equilibrium traffic flow given demand } q. \end{array}$$

† This is simply an economic definition. There is no assumption that the toll revenues are somehow injected back into the traffic system.

This MPEC can be written as an MPCC, assuming Q is specified by finitely many smooth equalities and inequalities, by using (2.2)-(2.3) to represent the equilibrium constraint.

(*b*) *Congestion control in telecommunication networks*

(i) *Rate flow equilibria for TCP*

Background: We re-work the standard model of Kelly *et al.* (1998) of the transmission control protocol, TCP, in a communication network like the internet, based on Kelly (2003, 2008). Again we have a network traffic of N nodes, a set of arcs $\mathcal{A}$ and a set of OD pairs $\mathcal{O}$. Path flows are intrinsic because, for a given OD pair $\kappa = kl$, the set of admissible routes $\mathcal{R}_\kappa$ from k to l, or routes that serve κ, is given by a routing protocol that is exogenous to TCP. In particular, neither the shortest path aspect of Wardrop equilibria nor remodelling path flows by link flows in its VI and MCP formulations can be readily applied to TCP.

Forward model of TCP equilibria: TCP is used at each node to determine the rate at which packets should be sent along outgoing routes. It responds to the level of unreliability of the network, due to congestion, which is manifested through loss of data packets. Let $\mathcal{R}$ denote the set of all routes used in the network. For each route r, TCP maintains a variable $\texttt{cwnd}_r > 0$ that is the size of the so-called congestion window. For every positive acknowledgement of a packet sent on r, $\texttt{cwnd}_r$ is increased by $a\texttt{cwnd}_r{}^\alpha$ and, for every lost packet, $\texttt{cwnd}_r$ is decreased by $b\texttt{cwnd}_r{}^\beta$. The parameters a, b, α and β are independent of the route, with $a, b > 0$ and $\alpha < \beta$. (In Kelly (2008), $a = 1$, $b = 1/2$, $\alpha = -1$ and $\beta = 1$.) Let p_r be the probability that a packet is lost on r, hence the average change in $\texttt{cwnd}_r$ per packet sent is $(1 - p_r)a\texttt{cwnd}_r{}^\alpha - p_r b\texttt{cwnd}_r{}^\beta$. When network flows are in equilibrium, the average change is zero and we derive

$$p_r \quad = \quad \left(1 + \frac{b}{a}(\texttt{cwnd}_r)^{\beta-\alpha}\right)^{-1}, \quad \forall r \in \mathcal{R}. \tag{2.4}$$

To understand the intrinsic meaning of p_r rather than the equilibrium identity (2.4), we look to probability of packet loss π_a on link a which, in turn, is related to the link capacity. The capacity u_a of link a is on the rate of flow, rather than the number of packets. Suppose the round trip time of sending out a packet on a route r and receiving a confirmation is roughly constant, denoted T_r, that is, congestion causes packet loss rather than delay. Then the flow rate on route r is roughly $z_r = \texttt{cwnd}_r/T_r$, and the total flow rate on link a is the sum of z_r over routes r containing a. The latter can be written Bz where $z = (z_r)_{r\in\mathcal{R}}$ and $B \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{R}|}$ is link-route incidence matrix: $B_{a,r}$ equals 1 if link a lies in r and 0 otherwise. Thus link capacity bounds are given by $Bz \le u$.

Returning to the likelihood of lost packets on link a, if the associated flow rate $\sum_{r\in\mathcal{R}} B_{r,a}z_r$ is strictly below u_a then π_a is deemed to be zero, while if the flow rate hits u_a then π_a may be positive. This is a complementarity relationship,

$$0 \le u - Bz \quad \perp \quad \pi \ge 0. \tag{2.5}$$

Moreover $1-p_r$ is the product of probabilities $1-\pi_a$ over all links a in r, which can be written $\Pi_{a\in\mathcal{A}}(1-\pi_a)^{B_{a,r}}$. We approximate the last product by $1 - \sum_{a\in\mathcal{A}} \pi_a B_{a,r}$,

assuming each π_a is small, thus $p_r = \sum_{a\in\mathcal{A}} \pi_a B_{a,r}$. We can express this more succinctly as

$$p = B^\top \pi. \tag{2.6}$$

MCP formulation of TCP equilibria: The conditions (2.4), (2.5) and (2.6) describe the rate flow equilibria of TCP but for one implicit assumption: z must be nonnegative. In fact we would expect the probability of packet loss on each route r to be less than one, and the associated flow rate, via (2.4), to be strictly positive. Nevertheless, for mathematical consistency we formulate an MCP that includes the constraint $z \geq 0$. To simplify (2.4) we simultaneously introduce the function $\phi : \mathbb{R}^{|\mathcal{R}|} \to \mathbb{R}^{|\mathcal{R}|}$,

$$\phi_r(z) = \left(1 + \frac{b}{a}(T_r z_r)^{\beta-\alpha})\right)^{-1}, \tag{2.7}$$

and eliminate p using (2.6), to form an MCP in which (2.5) is repeated:

$$\begin{array}{rcl} 0 \leq -\phi(z) + B^\top \pi & \perp & z \geq 0 \\ 0 \leq u - Bz & \perp & \pi \geq 0. \end{array} \tag{2.8}$$

This preserves the model motivation for routes r with $z_r > 0$.

VI formulation and existence of TCP equilibria: Let $Z = \{z \geq 0 : Bz \leq u\}$, a polyhedral convex set. Hence (Facchinei and Pang, 2003) the MCP (2.8) is equivalent to VI$(-\phi, Z)$. Note that the link-loss probability vector π is now implicit in the VI: it is the KKT multiplier or shadow cost of the link-capacity constraint $Bz \leq u$. We have assumed $\alpha < \beta$, above, which implies ϕ is continuous on Z. Also, given that B is the link-arc incidence matrix of a network, which we assume is connected, Z is a nonempty, compact convex set. Thus there exists a solution z of VI$(-\phi, Z)$, i.e., a flow rate equilibrium for TCP. We have also assumed $a, b > 0$, hence, from (2.7), each component function $\phi_r(z)$ depends only on z_r and is strictly decreasing in z_r. Thus, trivially, $-\phi$ is strictly monotone and uniqueness of the equilibrium follows.

In fact existence is usually shown (Kelly *et al.*, 1998; Kelly, 2003, 2008) via an optimization formulation of the equilibrium model, whose KKT conditions are (2.8). This optimisation formulation is the direct analog for TCP of the formulation of Wardrop traffic equilibrium by Beckmann *et al.* (1956). We make further comments on VI versus optimisation models in section (c).

(ii) *A design question for congestion control*

This and the next section speculate on the application of MPECs and MPCCs to congestion control in communication networks.

Suppose we want to choose to improve performance of the network. One performance metric is throughput $\sum_{r\in\mathcal{R}} z_r$. Another is average reliability which can be modelled as 1 less the probability of losing a packet anywhere in the network, the latter probability being $|\mathcal{R}|^{-1} \sum_{r\in\mathcal{R}} p_r$. There are several others studied in the literature, including fairness, stability, avoidance of trip time bias and so on (Kelly *et al.*, 1998; Kelly, 2003, 2008). These can be accounted in many ways, for example

by maximising a weighted sum of different performance objectives or including, in the constraints, lower bounds on performance metrics.

As a concrete example suppose we want to improve throughput without sacrificing reliability of any link. This question is studied in the recent work of Gu *et al.* (2007) for small buffer networks in which throughput and link speed is increasing, where an "evolutionary" version of TCP is proposed to prevent losses going to 100% as sending rates increase. Here however we model this as an MPCC. Let us set an upper bound $P \in (0,1)$ on probability of packet loss and insist that

$$P \geq p_r, \quad r \in \mathcal{R}. \tag{2.9}$$

An optimal design MPCC would be to maximise $\sum_{r\in\mathcal{R}} z_r$ over upper level variables a, b, α, β and lower level variables z, p, π subject to linear constraints on the former and the following requirements on the latter: z and π satisfy the complementarity constraints (2.8), p is given by (2.6) and satisfies the reliability constraints (2.9).

Looking beyond TCP, there are many layers in a communication network (Chiang *et al.*, 2007; Lin *et al.*, 2006), e.g., route selection is topic of its own. The design problem of optimising route selection and TCP parameters simultaneously, both of which effect the TCP equilibrium, is an MPEC that mixes combinatorial variables with continuous ones. This approach is unconvincing, however, because there is no single entity controls route selection and TCP parameters for all OD pairs. A decomposition approach that might have more realism is a game between OD pairs, where each agent chooses is own routes and TCP settings to maximise a measure of its performance. This falls into the category of "equilibrium program with equilibrium constraints" that are mentioned again in section 5.

(iii) *Inverse problems in transmission control*

We very briefly mention two possible inverse or calibration MPECs, both of which are attempts to identify user behaviour from a limited data sets.

The first is the direct analog of the origin-destination demand problem in traffic equilibrium problems in which we estimate some unknown OD rate flows in the knowledge of link flow rates and perhaps some particular route flow rates.

The second is to identify how aggressively users push packets through the network where users are represented as OD pairs. An aggressive (passive, respectively) user sets $a\texttt{cwnd}_r{}^{\alpha}$ large (small, resp.) relative to $b\texttt{cwnd}_r{}^{\beta}$. Suppose OD route flows are known. Then we can estimate user aggression via an MPEC in which the TCP parameters a, b, α, β for each OD pair, or a subset, are the upper level variables. These can be chosen differently for each OD pair subject to specified parameter constraints for that OD pair. The lower level variables would z and π, and would have to satisfy the MCP formulation (2.8) of TCP equilibrium. (It should be noted that unique existence of a TCP equilibrium does not rely on the TCP parameters being identical for all users.) The objective function would be the least squares difference between the OD route flows that are given as data and those which solve the lower level MCP.

(c) *Network design via convex optimisation or MPECs?*

We conclude this section on models with a brief discussion to illustrate some differences in the philosophy that underlies research in traffic equilibria and communication network equilibria.

Though VI and MCP models of equilibria are emphasized above, optimisation models are amongst the foundation stones of research on network equilibria. The seminal optimisation approaches of Beckmann *et al.* (1956) for Wardrop equilibria and Kelly *et al.* (1998) for TCP equilibria, which are brought together in Kelly and Voice (2005), underline this point. This view extends readily to the much more general setting of multi-layered setting of communication networks as surveyed by Chiang *et al.* (2007); Lin *et al.* (2006). In fact Chiang *et al.* (2007) use network utility maximisation (NUM) as a broad descriptor for the optimisation approach, and identify the "network as optimizer" and "protocol as a distributed solution". Part of the attraction of optimisation models, specifically convex optimisation so that global optimality is achievable, is that they seem easier to work with than equilibrium systems. Note that in the case of equilibria over networks, little modelling freedom seems to be lost by using an optimisation formulation. Theory for convex optimisation is deep and elegant, including duality which has been useful in understanding and designing decomposition schemes that account for the heterogeneous nature of network users and processes. This is amplified by the statement that "Industry adoption of 'layering as optimisation decomposition' has already started" in Chiang *et al.* (2007), where the example of basic TCP analysis (Kelly *et al.*, 1998) leading to TCP improvements (Jin *et al.*, 2004) is cited.

For the purposes of MPEC and MPCC modelling, the alternative to network utility maximisation would be to formulate equilibria via variational inequalities or MCPs (e.g., the KKT conditions of an optimisation problem). This highlights an obvious disjunction between the NUM and MPEC approaches. The first is founded on convex optimisation and the nexus it has, for network systems, with user equilibria. Indeed if an optimisation decomposition approach to network analysis and design yields difficult, say nonconvex, subproblems, then the given decomposition may have been "conducted in a wrong way" (Chiang *et al.*, 2007). MPECs by contrast are unavoidably nonconvex optimisation problems.

This disjunction is also related to differences in research directions taken in traffic networks versus communication networks. In particular, the computational and algorithmic sub-literature on bilevel optimisation (including MPECs) over traffic networks (for example, Chen and Florian (1998); Marcotte (1986); Smith (2004); Sumalee (2004a,b) and references therein) appears not to be paralleled in communication networks. Historically the role of civil engineering in road design and construction can be credited with the ongoing interest in traffic networks as static systems, i.e., the design variables will be fixed over much longer time periods than a driver would experience in a single journey. This is true whether considering which new roads to build and their capacity, or what traffic cordon and fee to set for a city like London. That is, MPECs have become important in applications for which static equilibria capture most of the system information needed for design.

By contrast, dynamics of network flows are central to communications networks where protocols like TCP are control mechanisms, so that issues like stability (Kelly *et al.*, 1998; Kelly, 2003; Kelly and Voice, 2005; Kelly, 2008) are very important.

There is a long history and growing research on dynamic traffic flow models of which we mention only a sample: Carey *et al.* (2003); Friesz *et al.* (1993); Kuwahara and Akamatsu (1997); Verhoef (2003). The application of MPEC modelling in dynamic network flows is minimal to the author's knowledge.

3. Lagrangian stationary conditions for MPCCs

We shift our focus from model-specific applications of MPCCs to analysis of general MPCCs.

For even the simplest MPCC feasible set $\{(x, y) \in \mathbb{R}^2 : 0 \leq x \perp y \geq 0\}$, the obvious NLP reformulation of the constraints as

$$0 \leq x,\ 0 \leq y,\ 0 = xy$$

is problematic. For instance an infinitesimal perturbation of $0 = xy$ to $-\epsilon = xy$ with $\epsilon > 0$ makes the constraints infeasible. This lack of constraint stability implies (Robinson, 1976) violation of a standard and somewhat weak NLP assumption, the Mangasarian-Fromovitz constraint qualification. This constraint qualification is equivalent, at a local minimizer, to existence and boundedness of KKT multipliers (Gauvin, 1977).

Thus there are doubts that MPCCs reformulated as NLPs will have KKT multipliers at local minima. This calls into question the viability of NLP algorithms which, in the main, are validated theoretically by showing convergence to stationary points (e.g., Nocedal and Wright (1999)). Nevertheless we shall see that under a condition of linear independence of active constraint gradients, which holds in some sense generically, we can work with KKT-type conditions that characterise stationarity.

Most of the ideas of this section are presented informally in section (a) in the context of a simple MPCC that is a nearest point problem in $\mathbb{R}^2$. Section (b) gives a more general Lagrangian treatment of stationarity of MPCCs. The brief section (c) connects the previous results to stationarity of an equivalent NLP reformulation of an MPCC.

A final introductory note is that although we do not study second-order optimality conditions here, their development for MPCCs also parallels that for NLPs. See Nocedal and Wright (1999) for second-order conditions in standard nonlinear programming, and, e.g., Luo *et al.* (1996); Ralph and Wright (2004); Leyffer *et al.* (2007) where second-order conditions are developed or used for MPCCs.

(*a*) *Simple examples*

Let α and β be positive scalars, $f(x, y) = (x-\alpha)^2+(y-\beta)^2$, which is the square of the 2-norm $\|(x, y) - (\alpha, \beta)\|_2$, and consider the MPCC over two variables,

$$\min_{(x,y)\in\mathbb{R}^2} f(x, y) \ \text{ subject to } \ 0 \leq x \perp y \geq 0. \tag{3.1}$$

This is the Euclidean projection problem of finding the nearest point to (α, β) in the feasible set $\mathcal{F} = \{(x, y) : 0 \leq x \perp y \geq 0\}$. It has two local minimizers $(\alpha, 0)$ and $(0, \beta)$, one for each piece of the feasible set, $\{(x, 0) : x \geq 0\}$ and $\{(0, y) : y \geq 0\}$.

The *MPCC-Lagrangian* uses multipliers Λ^x and Λ^y for each of the constraints $x \geq 0$ and $y \geq 0$, and is defined as $\mathcal{L}(x, y, \Lambda^x, \Lambda^y) = f(x, y) - x\Lambda^x - y\Lambda^y$. This is the usual Lagrangian of the relaxed problem $\min_{(x,y)\geq 0} f(x, y)$. We continue with the terminology of Scheel and Scholtes (2000). A feasible point (x, y) is *weakly stationary* if there exist multipliers $\Lambda^x \perp x$ and $\Lambda^y \perp y$ such that

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \nabla_{(x,y)}\mathcal{L}(x, y, \Lambda^x, \Lambda^y) = -2\begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \begin{pmatrix} \Lambda^x \\ \Lambda^y \end{pmatrix}.$$

The orthogonality conditions mean that multipliers have no effect unless their respective constraints are active, i.e., hold with equality. If we had additional standard constraints such as $g(x, y) \geq 0$, these would be handled in the traditional way by introducing a multiplier Λ^g of the same dimension as $g(x, y)$, subtracting the term $g(x, y)^\top \Lambda^g$ from the MPCC-Lagrangian, and requiring $0 \leq g(x) \perp \Lambda^g \geq 0$.

Weak stationarity for (3.1) holds trivially at the origin by taking $(\Lambda^x, \Lambda^y) = -2(\alpha, \beta)$. If we were minimizing over the relaxed feasible set $\{(x, y) : (x, y) \geq 0\}$, the KKT conditions at would further require $(\Lambda^x, \Lambda^y) \geq 0$, because if $\Lambda^x < 0$ (which is true here) then $\nabla f(0, 0)^\top (1, 0) = \Lambda^x < 0$, i.e., (1,0) is a feasible descent direction that takes us away from the active constraint $x = 0$. A similar story can be told if $\Lambda^y < 0$. A critical observation is that such descent directions, which are entirely due to NLP technology, are also feasible directions for the MPCC. This leads to our next definition.

A feasible point (x, y) of the above MPCC called *strongly stationary* if it is weakly stationary and, if both inequality constraints are active, then $\Lambda = (\Lambda^x, \Lambda^y) \geq 0$. For example, (0,0) is a weakly stationary but not strongly stationary since both multipliers are negative, while the local minimum $(\alpha, 0)$ strongly stationary (with $\Lambda^x = 0$ and $\Lambda^y = -\beta < 0$) as is $(0, \beta)$.

Strong stationarity, like the KKT conditions for a standard NLP, is simple enough to embed in an algorithm. To emphasize the simplicity of the strong stationary conditions, consider an MPCC that is defined with $x, y \in \mathbb{R}^m$, objective function $f(x, y)$, and constraints in the same form $0 \leq x \perp y \geq 0$. The feasible set is the union of 2^m polyhedral convex pieces, each of which is described by

$$\begin{array}{ccccc} x_I & = & 0_I & \leq & y_I \\ x_J & \geq & 0_J & = & y_J \end{array} \tag{3.2}$$

where I is any subset of $\{1, \ldots, m\}$, J is its complement, x_I is the subvector $(x_i)_{i\in I}$ etc. A natural way to check stationarity of the origin, which is feasible for each piece, is to check the KKT conditions of each linearly constrained problem: minimise $f(x, y)$ subject to (3.2). This corresponds to what is called *primal stationarity* (Luo *et al.*, 1996) or *B-stationarity* (Scheel and Scholtes, 2000). As there are 2^m pieces, the computational effort needed to check stationarity in each piece separately would be at least $O(2^m)$. This is unattractive if not completely impractical.

What is the gap between strong and B-stationarity? Even when all constraint functions are linear, a local solution which is B-stationary need not be strongly stationary, as the next example shows:

$$\begin{array}{ll} \min\limits_{(x,y)\in\mathbb{R}^2} & x - 2y \\ \text{subject to} & g(x, y) = x - y \geq 0 \\ & 0 \leq x \perp y \geq 0. \end{array} \tag{3.3}$$

The pieces of the feasible set are the ray $\{(x,0) : x \geq 0\}$ and the degenerate piece $\{(0,0)\}$, which is what is left of $\{(y,0) : y \geq 0\}$ after applying the inequality $g(x,y) \leq 0$. Obviously the origin is B-stationary. The weakly stationary conditions at (0,0) require solutions $\Lambda^g \geq 0, \Lambda^x, \Lambda^y$ of

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ -2 \end{pmatrix} - \Lambda^g \begin{pmatrix} 1 \\ -1 \end{pmatrix} - \begin{pmatrix} \Lambda^x \\ \Lambda^y \end{pmatrix}.$$

It is not possible to solve this with nonnegative multipliers because this would require, from first equation, $\Lambda^g = 1 - \Lambda^x \leq 1$ and, from the second, $\Lambda^g = 2 + \Lambda^y \geq 2$.

Nevertheless in the next section we shall justify the claim that most extrema of most MPCCs are strongly stationary.

(*b*) *MPCC stationary conditions and a linear independence constraint qualification*

We give formal definitions for an MPCC without equality constraints,

$$\begin{array}{lll} \min\limits_{z} & f(z) & \\ \text{subject to} & g(z) \geq 0 & \\ & 0 \leq r(z) \quad \perp \quad s(z) \geq 0 \end{array} \tag{3.4}$$

where, as above, $f : \mathbb{R}^n \to \mathbb{R}$, $r, s : \mathbb{R}^n \to \mathbb{R}^m$, and $g : R^n \to \mathbb{R}^{m_g}$ are smooth. Omitting equalities will ease notation without sacrificing generality.

For a feasible point z, the *MPCC-active set* is given by the active constraint indices

$$\begin{aligned} I_g(z) &= \{i : g_i(z) = 0\} \\ I_r(z) &= \{i : r_i(z) = 0\} \\ I_s(z) &= \{i : s_i(z) = 0\}. \end{aligned}$$

That is, the active constraints of the MPCC ignoring orthogonality. Denote by $g_{I_g(z)}(z)$ the subvector $\big(g_i(z)\big)_{i \in I_g(z)}$ etc.

Definition 3.1. *Let z be feasible for the MPCC* (3.4)*. We say z is* B-stationary† *or* primal stationary *if for each partition $I \cup J$ of $\{1, \ldots, m\}$ such that $I \supset I_r(z)$ and $J \supset I_s(z)$, z is stationary for*

$$\begin{array}{lllll} \min\limits_{z} & f(z) & & & \\ \text{subject to} & g(z) & \geq & 0 & \\ & r_I(z) & = & 0 & \leq \quad s_I(z) \\ & r_J(z) & \geq & 0 & = \quad s_J(z). \end{array} \tag{NLP(I)}$$

An easy observation in the context of this definition is that each feasible point near z lies in the feasible set of of one of the pieces (NLP(I)). This endows B-stationarity with a certain power, namely, if all functions defining the MPCC are linear, hence

† This is actually the definition for primal stationarity from Luo *et al.* (1996), but it is equivalent to B-stationarity defined by Scheel and Scholtes (2000). We use the latter term in preference because it seems to be more prevalent in the literature.

each (NLP(I)) is a linear program, then a feasible point is a local minimizer if and only if it is B-stationary. That is, B-stationarity is a "tight" local optimality condition up to first order. Regrettably, B-stationarity has a combinatorial nature as seen by the possibly large number of partitions $I \cup J$ in its definition. Another way of saying this is that the nonconvexity of MPCCs persists in their stationary conditions.

The MPCC-Lagrangian will be used to define KKT-like optimality conditions:

$$\mathcal{L}(z,\Lambda) \;=\; \mathcal{L}(z,\Lambda^g,\Lambda^r,\Lambda^s) \;=\; f(z) - g(z)^\top\Lambda^g - r(z)^\top\Lambda^r - s(z)^\top\Lambda^s \tag{3.5}$$

where $\Lambda = (\Lambda^g,\Lambda^r,\Lambda^s) \in \mathbb{R}^{m_g}\times\mathbb{R}^m\times\mathbb{R}^m$.

Definition 3.2. *Let z be feasible for the MPCC* (3.4).

1. *We say z is* weakly stationary *if there exists a multiplier tuple $\Lambda = (\Lambda^g,\Lambda^r,\Lambda^s)$, called the* MPCC-multiplier*, satisfying*

$$\begin{array}{rcll} \nabla_z\mathcal{L}(z,\Lambda) & = & 0 & \\ 0 \;\le\; g(z) & \perp & \Lambda \ge 0 & \\ 0 \;\le\; r_i(z) & \perp & \Lambda^r_i & i=1,\dots,m \\ 0 \;\le\; s_i(z) & \perp & \Lambda^s_i & i=1,\dots,m \end{array} \tag{3.6}$$

2. *z is called* strongly stationary *if it is weakly stationary with an MPCC-multiplier Λ such that*

$$\Lambda^r_i,\Lambda^s_i \ge 0, \quad i \in I_r(z)\cap I_s(z). \tag{3.7}$$

$I_r(z)\cap I_s(z)$ is called the *biactive* index set. If it is empty, weak and strong stationarity coincide.

Looking back at the examples of the previous section, MPCC-LICQ holds at each point of the feasible set $0 \le x \perp y \ge 0$, where x, y lie in $\mathbb{R}$ (or $\mathbb{R}^m$), however it fails at the origin for the MPCC (3.3) because the additional constraint $x - y \ge 0$ is active there. This helps to explain the fact that, as noted previously, $(0,0)$ is B-stationary but not strongly stationary for (3.3).

Suppose z is feasible for (3.4) and consider index sets I, J and a piece (NLP(I)) of the MPCC described in Definition 3.4. The Lagrangian of (NLP(I)) coincides with the MPCC-Lagrangian, so, denoting its multiplier tuple $\Lambda^I = (\Lambda^{gI},\Lambda^{rI},\Lambda^{sI})$, the KKT conditions of (NLP(I)) at z are

$$\begin{array}{rcl} \nabla_z\mathcal{L}(z,\Lambda^I) & = & 0 \\ 0 \;\le\; g(z) & \perp & \Lambda^{gI} \;\ge\; 0 \\ 0 \;\le\; r_J(z) & \perp & \Lambda^{rI}_J \;\ge\; 0 \\ 0 \;\le\; s_I(z) & \perp & \Lambda^{sI}_I \;\ge\; 0. \end{array}$$

It is easy to see that the strong stationarity conditions (3.6)-(3.7) imply nonnegativity of Λ^{rI}_J and Λ^{sI}_I, and hence that the above KKT conditions follow. For example, if $i \in J$ then either i is also biactive, in which case $\Lambda^{rI}_i \ge 0$, or $r_i(z) > 0 = \Lambda^{rI}_i$. Our mission in this section is to understand some situations in which B-stationarity implies strong stationarity.

We introduce a linear independence condition on the constraint gradients of the MPCC. This is based on the the standard NLP linear independence constraint qualification† (LICQ), often quoted to ensure unique existence of multipliers at a local solution of an NLP.

Definition 3.3. *Let z be feasible for the MPCC* (3.4)*. The* MPCC-LICQ *holds at z if the MPCC-active constraint gradients*

$$\{\nabla_z g_i(z) : i \in I_g(z)\} \cup \{\nabla_z r_i(z) : i \in I_r(z)\} \cup \{\nabla_z s_i(z) : i \in I_s(z)\}$$

are linearly independent.

The next result (Luo *et al.*, 1998; Scheel and Scholtes, 2000) says that B-stationarity and strong stationarity are equivalent at feasible points of an MPCC provided MPCC-LICQ holds. This result, which has a simple proof that we will sketch, brings local optimality of MPCCs within the realm of easy computation. It has become an essential prerequisite to the application of many NLP ideas to MPCCs.

Theorem 3.4. *Let z be a feasible point of the MPCC* (3.4) *at which MPCC-LICQ holds. Then, z is B-stationary if and only if it is strongly stationary.*

Proof. As already explained, strong-stationarity implies B-stationarity. For the converse, suppose z is (feasible and) B-stationary. The family of partitions $I \cup J$ given in Definition 3.1 corresponds to taking arbitrary subsets K of the biactive index set $I_r(z) \cap I_s(z)$, and assigning $I = I_r(z) \cup K$, $J = \{1 \ldots, m\} \setminus I$. We call these (I, J) pairs admissible.

Recall that the KKT multiplier $\Lambda^I = (\Lambda^{gI}, \Lambda^{rI}, \Lambda^{sI})$ of (NLP(I)) at z satisfies the Lagrangian equation $\nabla_z \mathcal{L}(z, \Lambda^I) = 0$. This multiplier is uniquely defined by due to linear independence of the active constraints, which follows from MPCC-LICQ. That is, the KKT multiplier corresponding to z for (NLP(I)) does not depend on I. Denote this unique multiplier $\Lambda = (\Lambda^g, \Lambda^r, \Lambda^s)$ and note that weak stationarity of z follows because $\nabla_z \mathcal{L}(z, \Lambda) = 0$ and the remaining conditions of (3.6) are given by the KKT conditions of (NLP(I)).

To see strong stationarity, look at the KKT conditions of (NLP(I)) which require $\Lambda^r_J \geq 0$ and $\Lambda^s_I \geq 0$. Each biactive index i lies in some I for some admissible (I, J). Likewise i lies in J' for some admissible (I', J'). This gives (3.7). □

A corollary, in light of previous comments, is that if the MPCC-LICQ holds at a local minimum z of an MPCC, then z is strongly stationary. Conversely if z is a feasible point of an MPCC that is strongly stationary, then it is a local minimum under the under the further conditions that the objective function is convex and the constraints functions are linear.

Scholtes and Stöhr (2001) provide an important companion to Theorem 3.4: the MPCC-LICQ holds at all points of generic MPCCs. For instance in the MPCC (3.3), perturbing the constraint function $g(x, y) = x - y$ to $g(x, y) = x - y + \epsilon$, for any nonzero ϵ, creates a feasible set which satisfies MPCC-LICQ at every feasible point. Taking Theorem 3.4 together with this genericity result demonstrates the earlier claim, at the end of section (a), that "most extrema of most MPCCs are strongly stationary".

† LICQ holds if the gradients of active constraints are linearly independent.

(c) An equivalent NLP and its stationary conditions

The bridge from NLPs to MPCCs provided above is indirect; it adapts NLP motivations and mathematical techniques to MPCCs. Nevertheless it extends to equivalent NLP reformulations of MPCCs, such as

$$\begin{array}{lrcl} \min\limits_{z} & f(z) & & \\ \text{subject to} & g(z) & \geq & 0 \\ & r(z), s(z) & \geq & 0 \\ & r(z)^\top s(z) & \leq & 0. \end{array} \tag{3.8}$$

This is equivalent to (3.4) in that z is a local minimum of the NLP if and only if z is local minimum of the MPCC. Recall the discussion at the start of this section, on instability of constraints of the type used in (3.8). Fortunately under MPCC-LICQ, in spite of this instability, KKT multipliers do exist at local solutions:

Proposition 3.5. *Let z be a feasible point of the MPCC* (3.4) *at which MPCC-LICQ holds. If z is a local minimum of the MPCC or its NLP reformulation* (3.8) *then it is stationary for the latter, i.e., has KKT multipliers.*

This is a variant of a result given by Fletcher *et al.* (2006). The proof is very straightforward but for brevity we refer the reader to the source.

Other NLP formulations could equally be considered with similar results. For example given $r(z), s(z) \geq 0$, we could write $r(z) \perp s(z)$ as $r(z)^\top s(z) = 0$, or $r_i(z)s_i(z)$ either ≤ 0 or $= 0$ for $i = 1, \ldots, m$. See Anitescu (2005b) regarding existence of KKT multipliers for the latter two componentwise NLP formulations.

4. Nonlinear programming approaches to solving MPCCs

We review four NLP approaches to finding local solutions, more accurately stationary points, of MPCCs. In section (a) we present two closely related methods that solve a sequence of standard NLPs, the regularisation and complementarity-penalty methods. In the sequential NLP framework, an MPCC is approximated by a family of NLPs, that depends on a single parameter $\epsilon > 0$, where each NLP has better posed constraints than the naive formulation (3.8). The idea is to solve a sequence of NLPs as $\epsilon \to 0$ in the hope of identifying a stationary point of the MPCC. We present the ideas behind the convergence analyses of the regularisation and complementarity-penalty method using Lagrangian analysis, from Scholtes (2001), Hu and Ralph (2004) and Ralph and Wright (2004). Indeed these papers owe a debt to earlier work of Fukushima and Pang (2000) on a sequential NLP method called smoothing.

In section (b) we sketch the interior-point method of Leyffer *et al.* (2007). This embeds traditional interior-point ideas, which execute only one linear system solve per iteration, into the complementarity-penalty framework. Finally, section (c) looks briefly at how the classical sequential quadratic programming or SQP method is applied directly to an NLP formulation like (3.8), see Fletcher and Leyffer (2004); Fletcher *et al.* (2006).

These three sections show a progression in the application of NLP ideas to solving MPCCs. Sequential NLP methods rely on a black-box NLP solver, whereas

the interior point approach adapts an NLP algorithm to handle MPCCs, and finally, SQP is a standard NLP method that is applied, again as black box, but now directly, to an equivalent NLP formulation.

We mention two algorithmic frameworks for MPCCs that will not be studied here. The first is specific to MPCCs, namely, piecewise or decomposition methods. The basic piecewise programming approach, which assumes that the kth iterate z^k is feasible for the MPCC, is to choose a piece NLP(I^k) of the MPCC (see Definition 3.1) whose feasible set contains z^k, and use this point to initiate one or more steps of an NLP method applied to NLP(I) that generates another feasible point z^{k+1} of NLP(I^k). If the NLP method also produces KKT multiplier estimates, and there exists a negative multiplier corresponding to a biactive index at z^{k+1}, the NLP piece might be updated by adding or dropping an index from I^k. References include Fukushima and Tseng (2002); Giallombardo and Ralph (2008); Luo *et al.* (1996, 1998); Scholtes and Stöhr (2001). The second algorithm framework is suited to an MPEC (1.1) in which the equilibrium constraints have a unique solution $y = y(x)$ for each relevant upper level vector x. Some general nonsmooth programming techniques can be used to minimize the implicit objective function $\psi(x) = f(x, y(x))$ assuming that the upper level constraints $(x, y) \in \Omega$ reduce to constraints only on x, and some sensitivity information (a "generalised Jacobian" matrix) for $y(x)$ is generated in addition to the lower level solution itself. The convergence theory identifies limit points of the sequence $\{x^k\}$ generated by an implicit programming method as a kind of weak stationary point called a "Clarke stationary" point. See Outrata *et al.* (1998) for the underpinnings of applications and algorithms suited to implicit programming.

Note that for practical purposes, Dirkse and Ferris (1999) provide a modelling framework, for the General Algebraic Modelling System (GAMS) and MATLAB environments, in for easy application of the regularisation, complementarity-penalty, smoothing and direct NLP methods to the same MPCC problem.

(*a*) *Regularisation and complementarity-penalty methods*

Regularisation, Scholtes (2001), also called relaxation, is based on the following NLP in which $\epsilon > 0$ is a parameter:

$$\begin{array}{lrcl} \min\limits_{x\in\mathbb{R}^n} & f(x) & & \\ \text{subject to} & g(x) & \geq & 0 \\ & r(x),\ s(x) & \geq & 0 \\ & r(x) * s(x) & \leq & \epsilon\mathbf{1} \end{array} \qquad \text{Reg}(\epsilon)$$

where all functions are twice continuously differentiable; $*$ is the Hadamard or componentwise product, so $r(x) * s(x)$ is the vector with ith component $r_i(x)s_i(x)$; and $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^m$. When $\epsilon = 0$, Reg(0) coincides with the NLP (3.8), hence is equivalent to the MPCC (3.4).

The idea is to find a local minimum x^k of Reg(ϵ_k) where $0 < \epsilon_k \to 0$. Now suppose x^* is a limit point of $\{x^k\}$. Then x^* is feasible for Reg(0) hence for the MPCC. We would like to know what further conditions are needed for x^* to be strongly stationary.

An easy preliminary result (Scholtes, 2001) is that if MPCC-LICQ holds at the x^*, then for small $\epsilon > 0$, the standard LICQ holds at feasible points of Reg(ϵ) near

x^*. Thus the regularised problem (with $\epsilon > 0$) has stability properties lacking in Reg(0). In particular a local minimum x^k of Reg(ϵ^k) will be stationary (Nocedal and Wright, 1999) if x^k is near x^* and ϵ_k is near 0.

The Lagrangian of Reg(ϵ_k) uses a multiplier tuple $\lambda = (\lambda^g, \lambda^r, \lambda^s, \lambda^{rs}) \in \mathbb{R}^{m_g+3m}$:

$$L_k^{\rm Reg}(x,\lambda) \;=\; f(x) - g(x)^\top\lambda^g - r(x)^\top\lambda^r - s(x)^\top\lambda^s + [r(x) * s(x) - \epsilon_k \mathbf{1}]^\top\lambda^{rs}$$

We link the stationary conditions of Reg(ϵ_k) to those of the MPCC through the gradient

$$\begin{aligned}
&\nabla_z L_k^{\rm Reg}(z,\lambda)\\
&\quad= \nabla f(z) - \nabla g(z)^\top\lambda^g - \nabla r(z)^\top[\lambda^r - s(z) * \lambda^{rs}] - \nabla s(z)^\top[\lambda^s - r(z) * \lambda^{rs}]\\
&\quad= \nabla_z \mathcal{L}(z, \Lambda(\lambda))
\end{aligned}$$

where $\mathcal{L}$ is the MPCC-Lagrangian, (3.5), and

$$\Lambda(\lambda) \;=\; (\lambda^g,\; \lambda^r - s(z) * \lambda^{rs},\; \lambda^s - r(z) * \lambda^{rs}).$$

Thus if λ^k is the KKT multiplier tuple for Reg(ϵ_k) at x^k and $\Lambda^k = \Lambda(\lambda^k)$, then

$$0 \;=\; \nabla_x L_k^{\rm Reg}(x^k, \lambda^k) \;=\; \nabla_x \mathcal{L}(x^k, \Lambda^k).$$

Let $\{x^k\}_{k\in\kappa}$ be a subsequence converging to x^*. Under MPCC-LICQ at x^*, the subsequence of approximate MPCC-multiplier tuples $\{\Lambda^k\}_{k\in\kappa}$ is itself convergent to a tuple $\Lambda^* = (\Lambda^{g*}, \Lambda^{r*}, \Lambda^{s*})$ by a basic stability property of a full rank matrix (corresponding, here, to the Jacobian at x^* of MPCC-active gradients); see Scholtes (2001) for details. Immediately we have $\nabla_x \mathcal{L}(x^*, \Lambda^*) = 0$ and indeed the remaining conditions of weak stationarity, see (3.6), follow by taking $k \in \kappa$, $k \to \infty$ in the KKT conditions of Reg(ϵ_k) at x^k.

In summary we give a result of Scholtes (2001):

Proposition 4.1. *Let x^k be stationary for* Reg(ϵ_k)*, where $0 < \epsilon_k \to 0$. If x^* is a limit point of $\{x^k\}$ at which MPCC-LICQ holds then it is weakly stationary for the MPCC (3.4).*

This result can be strengthened by adapting some ideas from Fukushima and Pang (2000), which addressed an alternative to regularisation called smoothing, that also solves a one-parameter family of NLP reformulation of MPCC. Assume in addition to z^k being stationary for Reg(ϵ_k) that it satisfies the *weak second-order necessary condition*: $H^k = \nabla^2_{zz} L_k^{\rm Reg}(z^k, \lambda^k)$ is positive semidefinite on the nullspace of constraint gradients with nonzero multipliers. That is $d^\top H^k d \geq 0$ for d satisfying

$$\begin{array}{lcll}
\nabla g_i(z^k)^\top d & = & 0, & \text{for } i \text{ such that } \lambda_i^{gk} > 0\\
\nabla r_i(z^k)^\top d & = & 0, & \text{for } i \text{ such that } \lambda_i^{rk} > 0\\
\nabla s_i(z^k)^\top d & = & 0, & \text{for } i \text{ such that } \lambda_i^{sk} > 0\\
\nabla [r_i(z^k) s_i(z^k)]^\top d & = & 0, & \text{for } i \text{ such that } \lambda_i^{rs\,k} > 0
\end{array}$$

where $\lambda^k = (\lambda^{gk}, \lambda^{rk}, \lambda^{sk}, \lambda^{rs\,k})$. For example, if x^k is a local minimum of Reg(ϵ_k) at which LICQ holds, then not only is it stationary, as discussed above, but the this second-order condition holds.

A final requirement is that the MPCC-multiplier $\Lambda^* = (\Lambda^{g*}, \Lambda^{r*}, \Lambda^{s*})$ satisfies *upper level strict complementarity*: both Λ_i^{r*} and Λ_i^{s*} are nonzero for $i \in I_g(z^*) \cap I_s(z^*)$. We quote from Scholtes (2001):

Theorem 4.2. *Let $\{z^k\}$, $\{\epsilon^k\}$ and z^* be as in Proposition 4.1, where the MPCC-LICQ holds at z^*, and z^* is weakly stationary with MPCC-multiplier Λ^*. If, in addition, each z^k satisfies the weak second-order necessary condition for* $\mathrm{Reg}(\epsilon_k)$ *and Λ^* satisfies upper level strict complementarity, then z^* is strongly stationary for the MPCC.*

The onus of the proof (Scholtes, 2001), which we omit, is to show that under the second-order condition on iterates z^k, the multipliers Λ^{r*} and Λ^{s*} are such that for biactive indices i, either $\Lambda_i^{r*}\lambda_i^{s*} = 0$ or Λ_i^{r*}, $\Lambda_i^{s*} > 0$. Hence upper level strict complementarity forces positivity of the biactive multipliers, i.e., strong stationarity holds.

Computational results in Scholtes (2001) indicate that regularisation can be more effective than direct application of a solver to the naive NLP formulation (3.8), while further analysis of regularisation can be found in Ralph and Wright (2004).

We make a quick comparison of regularisation with the complementarity-penalty approach, in which the troublesome constraint $r(z)^\top s(z) \leq 0$ in the NLP formulation (3.8) is omitted at the cost of adding a penalty term $\epsilon^{-1} r(z)^\top s(z)$ to the objective:

$$\begin{array}{lrcl} \min\limits_{z\in\mathbb{R}^n} & f(z) + \epsilon^{-1} r(z)^\top s(z) & & \\ \text{subject to} & g(z) & \geq & 0 \\ & r(z),\ s(z) & \geq & 0 \end{array} \qquad \mathrm{Pen}(\epsilon)$$

Below we mention some results from Hu and Ralph (2004); Ralph and Wright (2004). Further literature on penalty-complementarity methods for MPCC can be found in Anitescu (2005b,a); Anitescu *et al.* (2005); Huang *et al.* (2006).

As in regularisation, the idea is to find a local minimum or stationary point z^k of the penalty problem with $\epsilon = \epsilon_k > 0$, where $\epsilon_k \to 0$. Also, the usual LICQ holds for any feasible point z of the penalty problem that is near to a feasible point z^* of the MPCC at which MPCC-LICQ holds, thus we can expect local minima of $\mathrm{Pen}(\epsilon)$ to be stationary.

Next, suppose z^* is a limit point of the sequence $\{z^k\}$. Then z^* satisfies the constraints of the MPCC with the possible exception of orthogonality between $r(z^*)$ and $s(z^*)$. Thus feasibility of z^* must be assumed in order to discuss its stationarity properties.

Assuming feasibility of z^*, the convergence results of the complementarity-penalty method (Hu and Ralph, 2004) mirror those of the regularisation method given above. For example, Theorem 4.2 holds if the regularisation subproblem $\mathrm{Reg}(\epsilon_k)$ is replaced by $\mathrm{Pen}(\epsilon_k)$ for each k, and z^* is feasible for the MPCC. The proof technique (Hu and Ralph, 2004) is again Lagrangian.

A surprise, however, is that the complementarity-penalty method also exhibits a kind of "exactness", meaning that a solution of the penalty problem can be a solution of the original problem. What makes this surprising is that for NLPs that are reformulated via smooth penalty terms, as is done here, a stationary point of the penalty problem is typically not feasible (or stationary) for the original problem except as the penalty parameter goes to its limit, which is a theoretical rather than computational possibility. The next result due to Ralph and Wright (2004):

Proposition 4.3. *Assume z^* is a feasible point of the MPCC at which MPCC-LICQ holds, and that z^* is strongly stationary with MPCC-multiplier Λ^* that sat-*

isfies strict upper level complementarity. Then for a neighbourhood U of z^* and sufficiently small $\epsilon > 0$, each stationary point of Pen(ϵ) *in U is strongly stationary for the MPCC.*

This allows us to state an exact version of Theorem 4.2 for the complementarity-penalty method that allows for finite termination of the method at a strongly stationary point of the MPCC.

Theorem 4.4. *Let z^k be a stationary point of* Pen(ϵ_k) *where $0 < \epsilon_k \to 0$. Let z^* be the limit of a subsequence $\{z^k\}_{k\in\kappa}$ that is feasible for the MPCC, satisfies the MPCC-LICQ, and is weakly stationary with MPCC-multiplier Λ^*. If, in addition, each z^k satisfies the the weak second-order condition for* Pen(ϵ_k) *and Λ^* satisfies upper level strict complementarity then, for large enough $k \in \kappa$, each z^k is strongly stationary for the MPCC.*

Proposition 4.3 actually holds under a partial version of upper level strict complementarity (Ralph and Wright, 2004): for each biactive index i, at least one of Λ_i^{r*} and Λ_i^{s*} must be positive. The onus of the proof is to show that for sufficiently small $\epsilon > 0$, and a stationary point z^ϵ of Pen(ϵ) that is sufficiently close to z^*, that $r(z^\epsilon)^\top s(z^\epsilon) = 0$, that is, z^ϵ is feasible for MPCC.

We sketch the easy proof of complementarity at z^ϵ, where we suppose $0 < \epsilon \to 0$ and $z^\epsilon \to z^*$. By mimicking the proof of Proposition 4.1, it is easy to see that

$$\lambda_i^{r\epsilon} - \epsilon^{-1} s_i(z^\epsilon) \;\to\; \Lambda_i^{r*}$$

where $\lambda^{r\epsilon}$ is the KKT multiplier for Pen(ϵ) corresponding to $r(z^\epsilon) \geq 0$. We write this informally, as an approximate equality

$$\lambda_i^{r\epsilon} \;\approx\; \epsilon^{-1} s_i(z^\epsilon) + \Lambda_i^{r*}. \tag{4.1}$$

Suppose i is a biactive index for z^* and, for example, $\Lambda_i^{r*} > 0$. Since $s_i(z^\epsilon) \geq 0$,

$$\epsilon^{-1} s_i(z^\epsilon) + \Lambda_i^{r*} \geq \Lambda_i^{r*} > 0,$$

hence $\lambda_i^{r\epsilon} > 0$ for small $\epsilon > 0$ by (4.1). The KKT conditions of Pen(ϵ), which include complementarity of $\lambda_i^{r\epsilon}$ and $r_i(z^\epsilon)$, therefore force $r_i(z^\epsilon) = 0$. The case $\Lambda_i^{s*} > 0$ similarly implies $s_i(z^\epsilon) = 0$.

For an index i that is not biactive, it is possible for Λ_i^{r*} or Λ_i^{s*} to be negative and the above argument needs to be strengthened. If $\Lambda_i^{r*} < 0$, for example, then we must have $r_i(z^*) = 0 < s_i(z^*)$. That is, $s_i(z^\epsilon) > s_i(z^*)/2 > 0$ for small $\epsilon > 0$ and it follows from (4.1) that $\lambda_i^{r\epsilon}$ is positive for small $\epsilon > 0$. As above, this forces $r_i(z^\epsilon) = 0$. Likewise, the case $\Lambda_i^{s*} < 0$ yields $s_i(z^\epsilon) = 0$.

(*b*) *Penalty-interior approach*

The penalty-interior approach of Leyffer *et al.* (2007) is an interior-point method based on the complementarity-penalty version of sequential NLP. However it only solves one linear system per iteration, as do standard interior-point methods, rather than solving an NLP to optimality. The primal-dual interior-point framework is used here; see Wright (1997) for how this framework is used in standard optimisation.

Further literature on interior-point methods for MPCCs includes Benson *et al.* (2002); de Miguel *et al.* (2005); Luo *et al.* (1996); Raghunathan and Biegler (2005).

We simplify the setting of Leyffer *et al.* (2007) by assuming each constraint function g, r and s is linear (or affine). This will allow us to keep our notation simple by avoiding the introduction of nonnegative slack variables.†

The method uses two kinds penalty parameters, $\epsilon > 0$ to penalize deviation of $r(z)$ and $s(z)$ from orthogonality, and $\mu > 0$ to enforce nonnegativity of all inequality constraints via a log barrier penalty function. The penalty-interior formulation is

$$\min_{z\in\mathbb{R}^n} \quad f(z) + \epsilon^{-1} r(z)^T s(z) - \mu \sum_{i=1}^{m_g} \ln(g_i(z)) - \mu \sum_{i=1}^{m} [\ln(r_i(z)) + \ln(s_i(z))] \qquad \mathrm{Bar}(\epsilon, \mu)$$

This is an unconstrained problem because we have omitted equality constraints from the MPCC (3.4). However the domain of the objective function requires strict positivity of the vectors $r(x)$ and $s(x)$.

Sequential NLP provides some motivation, i.e., consider $\{z^k\}$ where each z^k solves $\mathrm{Bar}(\epsilon_k, \mu_k)$ for positive ϵ_k, μ_k, and $\epsilon_k, \mu_k \to 0$. The penalty-interior point method reduces the computational demand of sequential NLP by only approximately solving $\mathrm{Bar}(\epsilon_k, \mu_k)$, while increasing the solution accuracy as k increases. We might call this an approximate sequential NLP method.

At iteration k, given $\epsilon_k, \mu_k > 0$, the penalty-interior method uses two tolerances $\delta_k^{rs}, \delta_k > 0$ to bound solution inexactness:

$$\| \min\{r(z^k),\, s(z^k)\} \| \quad \leq \quad \delta_k^{rs} \tag{4.2}$$

$$\| \nabla_z L^{\mathrm{Bar}}(z^k; \epsilon_k, \mu_k) \| \quad \leq \quad \delta_k \tag{4.3}$$

where $L^{\mathrm{Bar}}(z; \epsilon, \mu)$ is the objective function of $\mathrm{Bar}(\epsilon, \mu)$, and the min operator in (4.2) is componentwise, with ith component equal to $\min\{r_i(z), s_i(z)\}$. This min operator is useful because for scalars α, β, the complementarity condition $0 \leq \alpha \perp \beta \geq 0$ holds if and only if $\min\{\alpha, \beta\} = 0$. The imposition of (4.2) is not unlike the regularisation method in that if $\delta_k^{rs} \to 0$, which will be assumed later, then limit points of the iteration sequence $\{z^k\}$ must be feasible for the MPCC. Note further that (4.3) corresponds to the inexact solution of the stationary conditions of $\mathrm{Bar}(\epsilon_k, \mu_k)$, and can easily be extended to handle equality constraints via first-order Lagrangian conditions.

There are many possibilities for updating the parameters ϵ_k, μ_k, δ_k^{rs} and δ_k. The most basic requirement is that all parameters approach zero as $k \to \infty$. Assuming this and that iterates satisfy (4.2)-(4.3) for each k, global convergence, i.e., properties of limit points of $\{z^k\}$, can be established with results and proofs that are similar to those of the regularisation method. See Leyffer *et al.* (2007) for details.

The "dynamic" version of the interior complementary-penalty method introduces two inner loops, one to decrease the complementarity-penalty parameter until (4.2) holds, the other to carry out more Newton steps on the barrier subproblem until (4.3) holds. We state a simple version of this.

† Strict nonnegativity, in interior-point methods, is much more conveniently handled if all inequalities are linear.

1. **Major step.** Given k, z^{k-1} at which g, r, s are positive, and positive parameters ϵ_k, μ_k, δ_k^{rs}, δ_k, initialise the inner loop: $z = z^{k-1}$, $\epsilon = \epsilon_k$.
2. **Inner loop.**
   (a) Find z^+ by applying one step of Newton's method to $\nabla_z L^{\rm Bar}(z; \epsilon, \mu_k) = 0$ at z.
   (b) If $\| \min\{r(z^+),\, s(z^+)\}\| > \delta_k^{rs}$, update $\epsilon = \epsilon/10$, GO TO 2(a).
   (c) Else, if $\|\nabla_z L^{\rm Bar}(z^+; \epsilon, \mu_k)\| > \delta_k$, update $z = z^+$, GO TO 2(a).
3. **Update.** Let $z^k = z^+$ and $\epsilon_{k+1} = \epsilon$; update μ_{k+1}, δ_{k+1} and δ_{k+1}^{rs}; update $k = k + 1$; GO TO 1.

The main work of the method is the Newton step in the inner loop, which requires a linear system solve:

$$z^+ \;=\; z - \nabla_{zz} L^{\rm Bar}(z; \epsilon, \mu_k)^{-1} \nabla_z L^{\rm Bar}(z; \epsilon, \mu_k).$$

This requires invertibility of the Hessian matrix $\nabla_{zz} L^{\rm Bar}(z; \epsilon, \mu_k)$, which can be assumed for small enough ϵ and μ_k under on a second-order sufficient condition at a minimum of the MPCC (Leyffer *et al.*, 2007). A robust numerical method would include many other considerations such as a procedure for modifying the matrix in the Newton step if it is not sufficiently well conditioned, and a line search to ensure g, r and s remain strictly positive at z^+. What is more relevant for us here is that this simple method is locally well defined, and well behaved, as outlined next.

We summarise, without giving details, a local convergence result of Leyffer *et al.* (2007). This summary relies, as does the above discussion, on linearity of the constraint functions although this is not required in the paper. Theorem 4.4 of Leyffer *et al.* (2007) assumes z^* is a strongly stationary point of the MPCC at which the LICQ and an appropriate second-order sufficient condition holds. It further assumes that the MPCC-multiplier Λ^* at z^* has nonzero components for every (MPCC)-active constraint. The theorem says, roughly, that if

- an iterate z^k is near enough to z^* with $g(z^k), r(z^k), s(z^k) > 0$, and $\epsilon_k > 0$ is sufficiently small,
- (4.2) and (4.3) hold,
- $\mu_k > 0$ is near zero and $\mu_{k+1} > 0$ converges to zero faster than $\|\nabla_z F(z^k; \epsilon_k, \mu_k)\|$, and
- the positive tolerances δ_k and δ_k^{rs} converge to zero neither too quickly nor too slowly relative to μ_{k+1} (which can be arranged in practice),

then, first, the inner loop in Step 2 of the interior-penalty method requires only one Newton step in order proceed to the next iteration via Step 3 (so $\epsilon_{k+1} = \epsilon_k$); and, second, $\{z^k\} \to z^*$ superlinearly: $\|z^{k+1} - z^*\|/\|z^k - z^*\| \to 0$.

The main difference in assumptions needed for this result and for the previous convergence results for the regularisation and penalty-complementarity methods is that here we need a strict complementarity assumption on all MPCC-multipliers. This allows the proof to make use of some standard results from the literature on

interior-point methods. The proof of the result also uses an exactness property of the complementarity-penalty method, Lemma 4.3 of the same paper (which is a corollary of Proposition 4.3 above). Exactness appears in the above convergence result through the update $\epsilon_{k+1} = \epsilon_k$, i.e., ϵ_k remains a fixed positive constant.

The algorithm's commendable numerical behaviour, of one linear solve per major iteration resulting in superlinear convergence of iterates, is often seen in practice according to the computational experience of Leyffer *et al.* (2007). While conditions for superlinear convergence of interior-point methods in solving convex programs are well understood, the fact that this is obtained for MPCCs can be credited to the exactness property of the penalty-complementarity formulation.

(*c*) *Sequential quadratic programming*

Recall the NLP formulation (3.8), which has some kind of constraint degeneracy. As expected (Jiang and Ralph (1999); Scholtes (2001) and others), and demonstrated using the `MINOS` and `CONSTR` nonlinear programming codes, applying NLP methods directly to (3.8) can lead to numerical difficulties. It was thus quite a surprise when Fletcher and Leyffer (2004) applied their sequential quadratic programming (SQP) code `SQP-Filter` to a suite of test MPCCs and found that the method usually converged superlinearly to a local solution. This was the first time superlinear convergence was observed for a standard NLP method applied directly to an equivalent NLP reformulation of an MPCC.

Fletcher *et al.* (2006) explain the local superlinear convergence of SQP applied to (3.8); see below for a motivating but incomplete summary. See also the related work of Anitescu (2005b). In practice, the application of NLP codes to MPCCs has progressed from small or medium size problems to more demanding applications such as large-scale nonlinear MPCC models of an electricity market network with 20,000 variables and 10,000 constraints (Chen *et al.*, 2006).

SQP is the version of Newton's method as it applies to constrained NLPs (Nocedal and Wright, 1999). At each iteration, it forms a quadratic approximation of the Lagrangian that it minimises over the linearised constraints. The Lagrangian of (3.8) is

$$L^{\mathrm{NLP}}(z,\lambda) \;=\; f(z) - g(z)^\top\lambda^g - r(z)^\top\lambda^r - g(z)^\top\lambda^g + r(z)^\top s(z)\lambda^{rs}$$

where λ is broken into subvectors $(\lambda^g, \lambda^r, \lambda^s, \lambda^{rs}) \in \mathbb{R}^{m_g} \times \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}$. Given the kth primal-dual iterate (z^k, λ^k), SQP finds a stationary point Δz and associated KKT multiplier λ^{k+1} of the quadratic program (QP)

$$\begin{array}{lrcl} \min\limits_{\Delta z} & \nabla_z L^{\mathrm{NLP}}(z^k,\lambda^k)\Delta z \,+\, \frac{1}{2}\Delta z^\top \nabla^2_{zz} L^{\mathrm{NLP}}(z^k,\lambda^k)\Delta z & & \\ \text{subject to} & g(z^k) + \Delta g & \geq & 0 \\ & r(z^k) + \Delta r,\; s(z^k) + \Delta s & \geq & 0 \\ & r(z^k)^\top s(z^k) \,+\, \Delta r^\top s(z^k) \,+\, r(z^k)^\top \Delta r & \geq & 0 \end{array}$$

where $\Delta g = \nabla g(z^k)\Delta z$, $\Delta r = \nabla r(z^k)\Delta z$ and $\Delta s = \nabla s(z^k)\Delta z$. SQP defines $z^{k+1} = z^k + \Delta z$ and repeats the process from the new primal-dual point.

In addition to the standing assumption that all functions f, g, r and s are twice continuously differentiable, we assume:

A1 z^* is a strongly stationary point.

A2 MPCC-LICQ holds at z^*.

A3 The MPCC-multiplier $\Lambda = (\Lambda^g, \Lambda^r, \Lambda^s)$ has $\Lambda^r_i, \Lambda^s_i > 0$ for $i \in I_r(z^*) \cap I_s(z^*)$.

A4 A strong second-order sufficient condition holds: $\nabla_{zz}\mathcal{L}(z^*, \Lambda)$ is positive definite with respect to the cone of directions d that satisfy $\nabla g_i(z^*)_i \top d \geq 0$ for $i \in I_g(z^*)$ with equality if $\Lambda^g_i > 0$, and similar inequalities and equalities for the active constraint gradients of r and s at z^*.

A5 The QP solver used at each iteration produces KKT multipliers whose nonzero components correspond to a set of linearly independent constraints.

Theorem 4.5. *Let A1-A5 hold. Suppose a primal-dual iterate (z^k, λ^k) is such that z^k is feasible for the NLP (3.8), z^k is near z^* and $\nabla_x L^{\rm NLP}(z^k, \lambda^k)$ is near 0. Then SQP applied to the NLP will generate z^{k+1} that is also feasible, such that $\{z^k\} \to z^*$ superlinearly.*

To give a brief motivation for the proof of this result, recall that the MPCC feasible set is patchwork of standard NLP feasible sets $\{z : r_I(z) = 0 \leq s_I(z),\ r_J(z) \geq 0 = s_J(z)\}$, where $I \cup J$ partition $\{1, \ldots, m\}$. If z^k is feasible for the MPCC, then it is feasible for an NLP piece, say NLP(I^k). The key to proving superlinear convergence is to show that if $(z^k,\ \nabla_x L^{\rm NLP}(z^k, \lambda^k))$ is near $(z^*, 0)$ then not only is z^* a local solution of NLP(I^k), but SQP applied to (3.8) at z^k gives the same next iterate z^{k+1} as it would when applied to NLP(I^k) at z^k. Superlinear convergence of $\{z^k\}$ thus follows by checking that the conditions for superlinear convergence of SQP (Nocedal and Wright, 1999) applied to NLP(I^k) are valid.

5. Beyond MPECs

We conclude the paper with brief discussions of two new and active research areas relating to network equilibria and MPECs that we have not been able to explore in this article.

(a) *Price of anarchy*

The "price of anarchy" in a multi-user system refers to gap between the maximum system utility that a central planner could hypothetically achieve, and the system utility associated with a user equilibrium. Or, as stated in the survey paper Roughgarden (2007) the "extent to which competition approximates cooperation". See also the monograph Roughgarden (2005). A less pejorative description of this gap would be the "cost of choice". The maximum system utility, achievable if either a central planner could force individuals to act according to instructions or all users cooperated to maximise total welfare, cannot be less than the utility at an equilibrium. Therefore the gap must be at least nonnegative.

Consider a family $\mathcal{E}$ of equilibrium problems over finitely many users, e.g., Wardrop traffic equilibria where users are link flows flows. For convenience suppose every equilibrium problem E in $\mathcal{E}$ has a unique solution. Given E and associated user actions that are feasible, the system utility is the sum of user utilities. For example, returning to traffic networks as presented in part (a) of section 2, a feasible

flow vector f and travel duration of $d_a(f)$ on each link $a \in \mathcal{A}$ yield system utility $\sum_{a\in\mathcal{A}} d_a(f)f_a$. We also assume all user's utilities are nonnegative on the feasible set of the equilibrium problem. By user utility of the problem E, denoted $U(E)$ we mean the system utility at equilibrium. By maximum utility of E, $M(E)$, we mean the supremum of E's system utility over all feasible user actions. The price of anarchy (Roughgarden, 2007) of E is defined mathematically as $U(E)/M(E)$, which is the relative gap between maximum and user utility, and the price of anarchy of $\mathcal{E}$ is $\inf_{E\in\mathcal{E}} U(E)/M(E)$. It may seem surprising that nontrivial bounds, some tight, on the price of anarchy for Wardrop equilibria over networks have been established for quite a general classes of networks (Roughgarden, 2007).

An MPEC might also be used to improve system utility but, unlike the utility maximisation problem, the upper level variables of an MPEC are generally quite limited, and only indirectly affect user behaviour. Let's discuss this in the context of toll design MPEC in traffic networks. A given traffic equilibrium problem depends parametrically on design variables which are the set of tolled arcs $\hat{\mathcal{A}} \subset \mathcal{A}$ and the toll level $\tau \geq 0$. Denote this problem by $E(\mathcal{A}, \tau)$ and the class of these problems over a given set of admissible designs by $\mathcal{E}$. The design parameters affect user utility but have no net effect on system utility. Thus the maximum system utility $M(E(\mathcal{A}, \tau))$ is independent of the design $(\mathcal{A}, \tau)$, hence we write it $M(\mathcal{E})$. Let E_0 be a reference equilibrium program corresponding to a particular design, say $\mathcal{A} = \emptyset$ or $\tau = 0$ — the untolled traffic network. Assume we can find a local maximum of the MPEC, by choosing $\mathcal{A}$ and τ, such that the system utility at the resulting Wardrop equilibrium exceeds $U(E_0)$. Write this local maximum value of the MPEC as $M^{\mathrm{MPEC}}(\mathcal{E})$ so that $U(E_0) \leq M^{\mathrm{MPEC}}(\mathcal{E}) \leq M(\mathcal{E})$. Thus, trivially,

$$\frac{U(\mathcal{E})}{M(\mathcal{E})} \leq \frac{U(E_0)}{M(\mathcal{E})} \leq \frac{U(E_0)}{M^{\mathrm{MPEC}}(\mathcal{E})} \leq 1.$$

If the price of anarchy $U(\mathcal{E})/M(\mathcal{E}) << 1$ then we would hope even local improvements via solving an MPEC would improve actual network performance. A theoretical issue is how to improve these bounds on $U(E_0)/M^{\mathrm{MPEC}}(\mathcal{E})$. Another is the efficiency of the MPEC formulation, for example, can we bound the ratio between the global MPEC utility value and the maximum achievable utility $M(\mathcal{E})$ by a number less than 1?

(*b*) *Equilibrium programs with equilibrium constraints*

Equilibrium programs with equilibrium constraints (EPECs) describe games in which some or all players face MPECs. Such an example in communication networks was identified in part (b)(ii) of section 2. There each OD pair could optimise both its selection of routes and TCP settings, to maximise its performance, subject to the constraint that network flow rates would be at a TCP equilibrium. In traffic networks, the issue of setting toll cordons around cities in regional clusters can be thought of as an EPEC. Each city wants to optimise its cordon and cordon fee, but its design will affect traffic flow in other cities via inter city and through traffic flows. A preliminary study by Hult (2006) suggests that using a different cordon design around each city can significantly reduce congestion even if the same toll level is applied within each design. This study, however, did not attempt to identify designs that solved a corresponding EPEC.

EPECs have become somewhat popular in the study of electricity markets, for example Berry *et al.* (1999); Hu and Ralph (2007); Yao *et al.* (2008). (See also Chen *et al.* (2006) who describe a large-scale MPEC, solved via NLP techniques, that models electricity and nitrous oxide emissions in the PJM network in the north east of the United States.) In such markets, pricing and dispatch are typically carried out by solving an optimisation problem over constraints that describe the electricity grid, including capacity constraints on links, that minimises system cost in order to match supply to demand. The system operator assembles the system cost as the sum of cost functions provided by each generator in the market. If each generator bids its true (marginal) cost of production function, then the pricing and dispatch problem would meet demand at minimum cost, therefore would maximise consumer welfare. However generators are not constrained to bid their true cost functions. Each generator therefore faces a bilevel program: what cost function should it bid to the system operator to maximise its profit? This can be modelled and solved numerically using MPCC technology (Hu and Ralph, 2007).

Although EPECs are attractive for modelling user behaviour in bilevel systems, they can be intractable. The difficulty is that the dependence of the lower level equilibrium solution on the upper level parameters, which is often nondifferentiable (piecewise smooth), can induce nonconvexity in each player's objective function. See Hu and Ralph (2007) for an electricity market EPEC over two players for which there no (pure strategy Nash) equilibrium. This pathology occurs although many features of the problem seem mathematically well behaved: each player has only one strategic (upper level) variable that lies in a closed and bounded interval (i.e., a nonempty, convex, compact set), a convex quadratic objective function over all variables, and lower level equilibrium system in the form of a linear complementarity problem that is uniquely solvable for each set up upper level variables. Nevertheless such pathological cases need not dominate a class of EPECs. Hu and Ralph (2007) give conditions under which equilibria are guaranteed to exist for some special cases in electricity markets. Even in the absence of existence theory, it is often possible to find equilibria numerically (Hu and Ralph, 2007; Yao *et al.*, 2008) by taking advantage of nonlinear programming techniques.

Acknowledgement. The author gratefully acknowledges the input of Frank Kelly, Gaurav Raina, Mike Smith and Agachai Sumalee on the literature of network equilibria, and Michael Ferris, Jong-Shi Pang and Stefan Scholtes on technical developments for MPECs.

References

Anitescu, M. 2005a. Global convergence of an elastic mode approach for a class of mathematical programs with complementarity constraints. *SIAM Journal on Optimization* **16** (1), 120–145,.

Anitescu, M. 2005b. On using the elastic mode in nonlinear programming approaches to mathematical programs with complementarity constraints. *SIAM Journal on Optimization* **15** (4), 1203–1236.

Anitescu, M., P. Tseng, and S. J. Wright. 2005. Elastic-mode algorithms for mathematical programs with equilibrium constraints: global convergence and stationarity properties. Tech. Rep. ANL/MCS P1242-0405, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA.

Beckmann, M., B. B. McGuire, and C. B. Winsten. 1956. *Studies in the Economics of Transportation*. Yale University Press, Cowles Commission Monograph, New Haven, CT.

Benson, H., D. F. Shanno, and R. J. Vanderbei. 2002. Interior-point methods for nonconvex nonlinear programming: Complementarity constraints. Tech. Rep. ORFE 02-02, Department of Operations Research and Financial Engineering, Princeton University, to appear in *Mathematical Programming, Series A*.

Berry, C. A., B. F. Hobbs, W. A. Meroney, R. P. O'Neill, and W. R. Stewart Jr. 1999. Understanding how market power can arise in network competition: a game theoretic approach. *Utilities Policy* **8**, 139–158.

Carey, M., Y. E. Ge, and M. McCartney. 2003. A whole-link travel-time model with desirable properties. *Transportation Science* **37**, 83–96.

Chen, Y. and M. Florian. 1998. Congested O-D trip demand adjustment problem: formulation and optimality conditions. In Migdalas, A., P. Pardalos, and P. Värbrand (eds.), *Multilevel Optimization: Algorithms, Complexity and Applications*. Kluwer Academic Publishers, pp. 1–22.

Chen, Y., B. F. Hobbs, S. Leyffer, and T. S. Munson. 2006. Leader-follower equilibria for electric power and NO_x allowances markets. *Computational Management Science* **3**, 307–330.

Chiang, M., S. H. Low, R. Calderbank, and J. C. Doyle. 2007. Layering as optimization decomposition: a mathematical theory of network architectures. *Proceedings of IEEE* **95**, 255–312.

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

de Miguel, A. V., M. Friedlander, F. Nogales, and S. Scholtes. 2005. An interior-point method for MPECS based on strictly feasible relaxations. *SIAM Journal on Optimization* **16** (22), 587–609.

Dempe, S. 2003. Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints. *Optimization* **52**, 333–359.

Dirkse, S. and M. C. Ferris. 1999. Modelling and solution environments for MPEC: GAMS and MATLAB. In Fukushima, M. and L. Qi (eds.), *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods*. Kluwer Academic Publishers, pp. 127–148.

Facchinei, F. and J. S. Pang. 2003. *Finite-dimensional variational inequalities and complementarity problems*. Springer Series in Operations Research. Springer, New York.

Fletcher, R. and S. Leyffer. 2004. Solving mathematical program with complementarity constraints as nonlinear programs. *Optimization Methods and Software* **19**, 15–40.

Fletcher, R., S. Leyffer, D. Ralph, and S. Scholtes. 2006. Local convergence of SQP methods for mathematical programs with equilibrium constraints. *SIAM Journal on Optimization* **17**, 259–286.

Friesz, T. L., D. Bernstein, T. E. Smith, R. L. Tobin, and B. W. Wie. 1993. A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research* **41**, 179–91.

Fukushima, M. and J. S. Pang. 2000. Convergence of a smoothing continuation method for mathematical programs with complementarity constraints. In Théra, M. and R. Tichatschke (eds.), *Ill-posed Variational Problems and Regularization Techniques*. Springer, pp. 99–110.

Fukushima, M. and P. Tseng. 2002. An implementable active-set algorithm for computing a b-stationary point of a mathematical program with linear complementarity constraints. *SIAM Journal on Optimization* **12**, 724–739.

Gauvin, J. 1977. A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Mathematical Progamming* **12**, 136–138.

Giallombardo, G. and D. Ralph. 2008. Multiplier convergence in trust-region methods with application to convergence of decomposition methods for MPECs. *Mathematical Programming* **112**, 335–369.

Gu, Y., D. Towsley, C. V. Hollot, and H. Zhang. 2007. Congestion control for small buffer high speed networks. *Proceedings of IEEE INFOCOM 2007* , 1037–1045.

Hu, J., J. E. Mitchell, J. S. Pang, K. P. Bennett, and G. Kunapuli. 2007. On the global solution of linear programs with linear complementarity constraints. Technical report, http://www.optimization-online.org/DB_FILE/2007/03/1608.pdf.

Hu, X. and D. Ralph. 2004. Convergence of a penalty method for mathematical programming with complementarity constraints. *Journal of Optimization Theory and Applications* **123**, 365–390.

Hu, X. and D. Ralph. 2007. Using EPECs to model bilevel games in restructured electricity markets with locational prices. *Operations Research* **55**, 809–827.

Huang, X. X., X. Q. Yang, and D. L. Zhu. 2006. A sequential smooth penalization approach to mathematical programs with complementarity constraints. *Numerical Functional Analysis and Optimization* **27**, 71–98.

Hult, E. E. 2006. Road pricing design. MPhil dissertation, Judge Business School, University of Cambridge.

Jiang, H. and D. Ralph. 1999. QPECgen, a MATLAB generator for mathematical programs with quadratic objectives and affine variational inequality constraints **13**, 25–59.

Jin, C., X. Wei, and S. H. Low. 2004. FAST TCP: motivation, architecture, algorithms, performance. *Proceedings of IEEE INFOCOM 2004* **4**, 2490–2501.

Kelly, F. 2003. Fairness and stability of end-to-end congestion control. *European Journal of Control* **9**, 159–176.

Kelly, F. 2006. Road pricing: addressing congestion, pollution and the financing of britain's roads. *Ignenia* **29**, 34–40.

Kelly, F. 2008. The mathematics of traffic in networks. In Gowers, W. and J. Barrow-Green (eds.), *The Princeton Companion to Mathematics*. Princeton University Press, to appear.

Kelly, F., A. Maulloo, and D. Tan. 1998. Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of Operational Research Society* **49**, 237–252.

Kelly, F. and T. Voice. 2005. Stability of end-to-end algorithms for joint routing and rate control. *Computer Communications Review* **35**, 5–12.

Kuwahara, M. and T. Akamatsu. 1997. Decomposition of the reactive dynamic assignments with queues for a many-to-many origin-destination pattern. *Transportation Research Part B: Methodological* **31**, ??

Leyffer, S., G. López-Calva, and J. Nocedal. 2007. Interior methods for mathematical programs with complementarity constraints. *SIAM Journal on Optimization* **17**, 52–77.

Lin, X., N. B. Shroff, and R. Srikant. 2006. A tutorial on cross-layer optimization in wireless networks. *IEEE Journal on Selected Areas in Communications* **24**, 1452–1463.

Luo, Z. Q., J. S. Pang, and D. Ralph. 1996. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press, Cambridge, UK.

Luo, Z. Q., J. S. Pang, and D. Ralph. 1998. Piecewise sequential quadratic programming for mathematical programs with nonlinear complementarity constraints. In Migdalas, A., P. Pardalos, and P. Värbrand (eds.), *Multilevel Optimization: Algorithms, Complexity and Applications*. Kluwer Academic Publishers, pp. 209–229.

Marcotte, P. 1986. Network design problem with congestion effects: a case of bilevel programming. *Mathematical Programming* **34**, 142–162.

Nocedal, J. and S. J. Wright. 1999. *Numerical Optimization*. Springer, New York.

Outrata, J. V., M. Kocvara, and J. Zowe. 1998. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints*. Kluwer Academic Publishers.

Patriksson, M. 1994. *The Traffic Assignment Problem: Models and Methods*. VSP, Ultrecht, The Netherlands.

Raghunathan, A. U. and L. T. Biegler. 2005. Interior point methods for mathematical programs with complementarity constraints. *SIAM Journal on Optimization* **15** (3), 720–750.

Ralph, D. and S. J. Wright. 2004. Some properties of regularization and penalization schemes for MPECs. *Optimization Methods and Software* **19** (5), 527–556.

Robinson, S. M. 1976. Stability theory for systems of inequalities, Part II: Differentiable nonlinear systems. *SIAM Journal on Numerical Analysis* **13**, 497–513.

Roughgarden, T. 2005. *Selfish Routing and the Price of Anarchy*. MIT Press, Boston, USA.

Roughgarden, T. 2007. Selfish routing and the price of anarchy. *Optima* **74**, 1–14.

Scheel, M. and S. Scholtes. 2000. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity. *Mathematics of Operations Research* **25**, 1–22.

Scholtes, S. 2001. Convergence properties of a regularization scheme for mathematical programs with complementarity constraints. *SIAM Journal on Optimization* **11**, 918–936.

Scholtes, S. and M. Stöhr. 2001. How stringent is the linear independence assumption for mathematical programs with stationarity constraints? *Mathematics of Operations Research* **21**, 851–863.

Sheffi, Y. 1985. *Urban Transportation Networks*. Prentice-Hall Inc.

Smith, M. 1979. The existence, uniqueness and stability of traffic equilibrium. *Transportation Research* **13B**, 295–394.

Smith, M. 2004. Bilevel optimisation of prices in a variey of transportation models. In Mahmassani, H. (ed.), *Proceedings of the Sixteenth Inter- national Symposium on Transportation and Traffic Theory*. University of Maryland, pp. 1–21.

Sumalee, A. 2004a. *Optimal Road Pricing Scheme Design*. Ph.D. Dissertation, Institute for Transport Studies. University of Leeds.

Sumalee, A. 2004b. Optimal road user charging cordon design: A heuristic optimization approach. *Computer Aided Design in Civil and Infrastructure Engineering* **19**, 377–392.

Verhoef, E. 2003. Inside the queue: hypercongestion and road pricing in a continuous time-continuous place model of traffic congestion. *Journal of Urban Economics* **54**, 531–64.

Wardrop, J. G. 1952. *Some theoretical aspects of road traffic research*.

Wright, S. J. 1997. *Primal-Dual Interior-Point Methods*. SIAM, Philadelphia, USA.

Yao, J., I. Adler, and S. S. Oren. 2008. Modeling and computing two-settlement oligopolistic equilibrium in a congested electricity network. *Operations Research*, to appear.