Approximate Bilevel Programming via Pareto Optimization for Imputation and Control of Optimization and Equilibrium models

Jérôme Thai[1] and Rim Hariss[2] and Alexandre Bayen[3]

***Abstract*— We consider the problem of imputing the function that describes an optimization or equilibrium process from noisy partial observations of nearly optimal (possibly non-cooperative) decisions. We generalize existing inverse optimization and variational inequality problems to construct a novel class of multi-objective optimization problems: approximate bilevel programs. In this class, the "ill" nature of the complementary condition prevalent in bilevel programming is avoided, and residual functions commonly used for the design and analysis of iterative procedures, are a powerful tool to study approximate solutions to optimization and equilibrium problems. In particular, we show that duality gaps provide stronger bounds than ℓ_p norms of KKT residuals. The weighted criterion method is in some sense equivalent to existing formulations in the case of full observations. Our novel approach allows to solve bilevel and inverse problems under a unifying framework, via block coordinate descent, and is demonstrated on 1) consumer utility estimation and pricing and 2) latency inference in the road network of Los Angeles.**

I. Introduction

Optimization [4] and equilibrium [7] problems have a wide range of applications in, *e.g.*, economics, engineering, statistics, finance. In many scenarios, the outputs are easily observable, while we do not have access to the function meant to describe the process. Iyengar and Kang [8] and Keshavarz, Wang, and Boyd [10] focused on the *inverse convex optimization* (inverse CO) of imputing a convex objective from full observations of nearly optimal decisions. Bertsimas et al. [3] recently considered the *inverse variational inequality* problem of imputing the function that describes the *Variational Inequality* (VI) from full observations of approximate equilibria. The works outlined above present many applications: consumer utility estimation, latency inference in traffic networks, value function estimation in control etc. In general, *inverse problems* have been studied quite extensively; see [8], [10], [3] for a survey.

In the present work, we focus on the fact that data suffers from *missing values* mainly due to experimental limitations and we extend our previous work [16] with a rigorous mathematical framework for our formulation, combining ideas from computational mathematics, inverse, bilevel, and *multi-objective* (MO) or Pareto programming.

[1]Jérôme Thai is with the department of Electrical Engineering and Computer Sciences, University of California at Berkeley. jerome.thai@berkeley.edu

[2]Rim Hariss is with the Operations Research Center, Massachusetts Institute of Technology. rhariss@mit.edu

[3] Alexandre Bayen is with the department of Electrical Engineering and Computer Sciences, and the department of Civil and Environmental Engineering, University of California at Berkeley. bayen@berkeley.edu

We first observe that existing formulations of inverse CO and VI problems fall into two categories: (i) a *bilevel form* [8], [16] in which the function describing the process is chosen such that the induced optimal decision minimizes the distance to the observations, (ii) a *residual form* [10], [3] in which the function is imputed such that that data points are the closest to being optimal. The former allows more control on the fit to data points, while the latter avoids the 'ill' nature of complementary constraints; see [9], [11]. Hence, we propose a *combined bilevel-residual form* in the form of a MO problem, called *approximate bilevel program*, which leverages the benefits of both formulations.

Our novel formulation relies on the notions of *approximate solutions* to VI and CO problems which are formalized by using specific residual functions typically used for the design and analysis of iterative methods (*e.g.* stopping rules or certificates of suboptimality) to solve VIs [7, §10] and CO problems [6, §6],[4, §9]. In our work, and similarly to [10], [3], residuals are used to design merit functions which measure the agreement of the fitting CO and VI models to the observations. In particular, we find that duality gaps provide stronger bounds than any ℓ_p-norm of KKT residuals. These theoretical insights are critical for the choice of appropriate residuals to solve inverse problems.

Finally, our MO formulation allows better control by allowing to express preferences between Pareto optimal points; see [12]: in structural estimation, *theoretical and experimental* results suggest that more weight should be assigned on the fit to data points whereas in control, more weight should be put on the residual function, especially when the model is a good approximation of the reality.

In contrast to the MPEC literature which focused on developing specialized algorithms [11], [5], we apply the block coordinate descent (BCD) algorithm proposed in our previous work [16] for out numerical experiments. Since the proposed reformulation is in general convex in each block of variables, except for a concave constraint that is relaxed, we apply CO to the block updates using high quality CO solvers.

II. Motivating example: traffic assignment

Since CO and VI models have numerous applications; see [4], [7], [15], the potential of inverse VI and CO problems is huge. We present two applications.

We give a summary of the traffic assignment model; see [13, §2.2.2], [16, §2] for more details. We consider a network $(\mathcal{N}, \mathcal{A})$ with $\mathcal{N}$ the node set and $\mathcal{A}$ the set of directed arcs. Given a set of commodities $\mathcal{C} \subseteq \mathcal{N} \times \mathcal{N}$, a flow of demand

rate λ_k must be routed from s_k to t_k for each commodity $k = (s_k, t_k) \in \mathcal{C}$. The k-th *commodity flow vector* is denoted $\mathbf{x}^k = (x_a^k)_{a\in\mathcal{A}} \in \mathbb{R}_+^{\mathcal{A}}$. For each arc $a \in \mathcal{A}$, we are also given a *continuous positive nondecreasing* delay (or latency) function $s_a : \mathbb{R}_+^{\mathcal{A}} \to \mathbb{R}$ depending on the *aggregate flow* vector $\mathbf{v} = \sum_k \mathbf{x}^k \in \mathbb{R}_+^{\mathcal{A}}$. Beckmann et al. [2] considered the *separable case* in which $s_a(\cdot)$ is only function of the aggregate flow $v_a = \sum_k x_a^k$ on arc a, and proved the *User Equilibrium* (UE) (defined by Wardrop [17]) exists and is solution to the program:

$$\min_{\mathbf{x}} z(\mathbf{Z}\mathbf{x}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b},\, \mathbf{x} \succeq 0 \tag{1}$$

where $(\mathbf{x}^k)_{k\in\mathcal{C}}$ are stacked in an overall flow vector $\mathbf{x} \in \mathbb{R}^{|\mathcal{C}||\mathcal{A}|}$, $\mathbf{Z} \in \{0,1\}^{|\mathcal{A}|\times|\mathcal{C}||\mathcal{A}|}$ maps $\mathbf{x}$ to $\mathbf{v}$, *i.e.* $\mathbf{v} = \mathbf{Z}\mathbf{x} = \sum_k \mathbf{x}_k$, and $f : \mathbb{R}_+^{\mathcal{A}} \to \mathbb{R}$ is the *Beckmann function* on $\mathbf{v}$:

$$z(\mathbf{v}) = \sum_{a\in\mathcal{A}} \int_0^{v_a} s_a(u)du \tag{2}$$

However, delay functions s_a are in general difficult to estimate while aggregate flows v_a are easily measurable, but only on a small subset of arcs in the network, due to the cost of deploying and maintaining a sensing infrastructure in a large urban network.

III. Motivating example: consumer behavior

We also consider an *oligopoly* in which n firms produce n products indexed by $i = 1, \cdots, n$ (one for each firm) with prices $\mathbf{p} = (p_i)_{i=1}^n$. We suppose that the consumer purchases a quantity x_i of product i in order to maximize a nondecreasing and concave utility function $U(\mathbf{x})$ minus the price paid $\mathbf{p}^T\mathbf{x}$, where $\mathbf{x} = (x_i)_{i=1}^n$ is the overall demand:

$$\min\, \mathbf{p}^T\mathbf{x} - U(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \succeq 0 \tag{3}$$

However, the utility $U : \mathbb{R}^n \to \mathbb{R}$ is not known in practice, and a method for imputing U based on N observations of pairs $(\mathbf{p}^j, \mathbf{x}^j)$, $j = 1, \cdots, N$ of prices and associated demands has been proposed in [10]. The imputed utility U is then used by company producing i to set a price p_i in order to achieve a target consumer demand x_i^{des} in its product.

In oligopolies, the price of each product is publicly available and each firm has in general perfect knowledge of its own demand x_i, however it may only have partial information on other demands.

IV. Variational inequality, convex optimization

We recall fundamental results in VI and CO theory. From our assumptions on the delay functions s_a (positivity, continuity, monotonicity), (1) is a convex program. Concavity of U also implies convexity of problem (3). Both problems have general form, denoted OP($\mathcal{K}, f$):

$$\min\, f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{K} \tag{4}$$

where $\mathcal{K} \subseteq \mathbb{R}^n$ is a convex set, $f : \mathcal{K} \to \mathbb{R}$ a convex function, and n the dimensionality. OP($\mathcal{K}, f$) is a *convex optimization* (CO) problem. From the optimality conditions [4, §4.2.3]:

Theorem 1. *With f differentiable and ∇f its gradient, a feasible point $\mathbf{x}^\star \in \mathcal{K}$ solves OP($\mathcal{K}, f$) if and only if*

$$\nabla f(\mathbf{x}^\star)^T(\mathbf{u} - \mathbf{x}^\star) \geq 0,\, \forall\, \mathbf{u} \in \mathcal{K} \tag{5}$$

The VI problem can be seen as a generalization of (5) where ∇f is replaced by a general map F. Given a set $\mathcal{K} \subseteq \mathbb{R}^n$ and a map $F : \mathcal{K} \to \mathbb{R}^n$, the VI problem, denoted VI($\mathcal{K}, F$), consists in finding $\mathbf{x} \in \mathcal{K}$ such that:

$$F(\mathbf{x})^T(\mathbf{u} - \mathbf{x}) \geq 0,\, \forall\, \mathbf{u} \in \mathcal{K} \tag{VI}$$

In the remainder of the article, we suppose that $\mathcal{K}$ is a polyhedron (such as in our two application):

$$\mathcal{K} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b},\, \mathbf{x} \succeq 0\} \tag{6}$$

This assumption allows different characterizations of solutions to VI($\mathcal{K}, F$). As in [1], we define the *primal-dual system* associated to the linear program $\min_{u\in\mathcal{K}} F(\mathbf{x})^T\mathbf{u}$:

$$F(\mathbf{x})^T\mathbf{x} = \mathbf{b}^T\mathbf{y}, \quad \mathbf{A}^T\mathbf{y} \preceq F(\mathbf{x}), \quad \mathbf{A}\mathbf{x} = \mathbf{b},\, \mathbf{x} \succeq 0 \tag{7}$$

From LP strong duality, we have the following result:

Theorem 2. *Let $\mathcal{K}$ polyhedral given by (6). Then $\mathbf{x}$ solves VI($\mathcal{K}, F$) if and only if there exists $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y})$ satisfies the primal-dual system (7).*

The proof is given in [1, Th. 1]. A direct implication is:

Corollary 1. *Let $\mathcal{K}$ polyhedral given by (6) and f differentiable convex. Then $\mathbf{x} \in \mathcal{K}$ solves OP($\mathcal{K}, f$) if and only if there exists $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y})$ satisfies the primal-dual system (7) with $F = \nabla f$.*

A related system is the *Karush-Kuhn-Tucker* system of a VI:

$$F(\mathbf{x}) = \mathbf{A}^T\mathbf{y} + \boldsymbol{\pi},\ \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \succeq 0,\ \boldsymbol{\pi} \succeq 0,\ \mathbf{x}^T\boldsymbol{\pi} = 0 \tag{8}$$

The following result is from [7, §1.2.1]:

Theorem 3. *Let $\mathcal{K}$ polyhedral given by (6). Then $\mathbf{x}$ solves VI($\mathcal{K}, F$) if and only if there exists $\mathbf{y}, \boldsymbol{\pi}$ such that $(\mathbf{x}, \mathbf{y}, \boldsymbol{\pi})$ satisfies the KKT system (8).*

In convex optimization, the above result is known as the well-known KKT optimality conditions [4, §5.5.3].

Corollary 2. *Let $\mathcal{K}$ polyhedral given by (6) and f differentiable convex. Then a point $\mathbf{x} \in \mathcal{K}$ solves OP($\mathcal{K}, f$) if and only if there exists $\mathbf{y}, \boldsymbol{\pi}$ such that $(\mathbf{x}, \mathbf{y}, \boldsymbol{\pi})$ satisfies the KKT system (8) with $F = \nabla f$.*

We say that $\mathbf{x}$ is *primal feasible* if $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{x} \succeq 0$ and $(\mathbf{x}, \mathbf{y})$ is said to be *dual feasible* if $\mathbf{A}^T\mathbf{y} \preceq F(\mathbf{x})$. In the reverse OP($\mathcal{K}, f$), f is sought such that data points approximately satisfy the KKT system with $F = \nabla f$ [10], and in the reverse VI($\mathcal{K}, F$), F is sought such that data points approximately verify the primal-dual system [3].

V. Function imputation via Pareto optimization

Building on previous works [8], [10], [3], [16], we develop a novel class *multi-objective* (MO) programs for solving the inverse CO and VI problems.

Since the reverse OP($\mathcal{K}, f$) can be seen as finding a map $F = \nabla f$ and reconstructing f from F, we will refer to both VI($\mathcal{K}, F$) and OP($\mathcal{K}, f$) as a *mathematical problem* MP($\mathcal{K}, F$) over the feasible set $\mathcal{K}$. We first define a notion of approximate solutions:

Definition 1. *Given nonnegative functions r_{PD}, r_{KKT} such that*

$$r_{PD}(\mathbf{x}, \mathbf{y}) = 0 \iff \text{(7) holds at } (\mathbf{x}, \mathbf{y}) \tag{9}$$
$$r_{KKT}(\mathbf{x}, \mathbf{y}, \boldsymbol{\pi}) = 0 \iff \text{(8) holds at } (\mathbf{x}, \mathbf{y}, \boldsymbol{\pi}) \tag{10}$$

a point $\mathbf{x}$ is an ϵ-approximate solution to MP($\mathcal{K}, F$) under r_{PD} if there exists $\mathbf{y}$ such that $r_{PD}(\mathbf{x}, \mathbf{y}) \leq \epsilon$ (resp. under r_{KKT} if there exists $(\mathbf{y}, \boldsymbol{\pi})$ such that $r_{KKT}(\mathbf{x}, \mathbf{y}, \boldsymbol{\pi}) \leq \epsilon$).

The functions $r_{\text{PD}}, r_{\text{KKT}}$ are called *residual functions* associated to the primal-dual and KKT systems respectively. We are given N pairs $(\mathbf{q}^j, \mathbf{z}^j)$ of parameters and observations of nearly optimal decisions corresponding to different configurations of the system. The parameters $\mathbf{q}^j$ are triplets $(\mathbf{A}^j, \mathbf{b}^j)$ such that $\mathbf{A}^j, \mathbf{b}^j$ define a polyhedron $\mathcal{K}^j = \{\mathbf{x} \,|\, \mathbf{A}^j\mathbf{x} = \mathbf{b}^j\}$. We seek a map F and approximate solutions $\mathbf{x}^j$ to MP($\mathcal{K}^j, F$) that agree with the observations. The inverse problem can be formulated in *residual form*:

$$\begin{array}{ll} \min\limits_{F,\mathbf{x},\mathbf{y}} & \sum_j r_{\text{PD}}(\mathbf{x}^j, \mathbf{y}^j) \\ \text{s.t.} & \mathbf{H}\mathbf{x}^j = \mathbf{z}^j,\ \mathbf{A}^{jT}\mathbf{y}^j \preceq F(\mathbf{x}^j),\ \forall\, j \\ & F \in \mathcal{F} \end{array} \tag{11}$$

$$\begin{array}{ll} \min\limits_{F,\mathbf{x},\mathbf{y},\boldsymbol{\pi}} & \sum_j r_{\text{KKT}}(\mathbf{x}^j, \mathbf{y}^j, \boldsymbol{\pi}^j) \\ \text{s.t.} & \mathbf{H}\mathbf{x}^j = \mathbf{z}^j,\ \mathbf{A}^{jT}\mathbf{y}^j \preceq F(\mathbf{x}^j),\ \boldsymbol{\pi}^j \succeq 0,\ \forall\, j \\ & F \in \mathcal{F} \end{array} \tag{12}$$

where $\mathcal{F}$ is the set of feasible maps. Observe that with full observations, *i.e.* with $\mathbf{H}$ the identity, (12) and (11) are equivalent to the formulations in [10] and [3] respectively. As we will see in the next section, imposing primal-dual feasibility on the pairs $(\mathbf{x}^j, \mathbf{y}^j)$, *i.e.* having $\mathbf{A}^j\mathbf{x}^j = \mathbf{b}^j$, $(\mathbf{A}^j)^T\mathbf{y}^j \preceq F(\mathbf{x}^j)$, $j = 1, \cdots, N$ allows a rigorous study of the residuals r_{PD}, r_{KKT}. However, since observations are noisy and mathematical models are approximations of the reality, the system $\mathbf{A}^j\mathbf{x}^j = \mathbf{b}^j$, $\mathbf{H}\mathbf{x}^j = \mathbf{z}^j$ may be infeasible, especially when we have full observations and would expect $\mathbf{A}^j\mathbf{z}^j = \mathbf{b}^j$ to hold. Hence, we relax $\mathbf{H}\mathbf{x}^j = \mathbf{z}^j$ and impose primal-dual feasibility, where ϕ is some penalty function:

$$\begin{array}{ll} \min\limits_{F,\mathbf{x},\mathbf{y}} & [\,\Sigma_j r_{\text{PD}}(\mathbf{x}^j, \mathbf{y}^j),\ \Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j)\,]^T \\ \text{s.t.} & \mathbf{A}^j\mathbf{x}^j = \mathbf{b}^j,\ \mathbf{x}^j \succeq 0 \quad \forall\, j \\ & \mathbf{A}^{jT}\mathbf{y}^j \preceq F(\mathbf{x}^j) \qquad \forall\, j \\ & F \in \mathcal{F} \end{array} \tag{13}$$

$$\begin{array}{ll} \min\limits_{F,\mathbf{x},\mathbf{y},\boldsymbol{\pi}} & [\,\Sigma_j r_{\text{KKT}}(\mathbf{x}^j, \mathbf{y}^j, \boldsymbol{\pi}^j),\ \Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j)\,]^T \\ \text{s.t.} & \mathbf{A}^j\mathbf{x}^j = \mathbf{b}^j,\ \mathbf{x}^j \succeq 0 \qquad \forall\, j \\ & \mathbf{A}^{jT}\mathbf{y}^j \preceq F(\mathbf{x}^j),\ \boldsymbol{\pi}^j \succeq 0 \quad \forall\, j \\ & F \in \mathcal{F} \end{array} \tag{14}$$

In the our MO formulations above, we are minimizing pairs of objective functions, and we also allow different choices of penalties on the observation residuals $\mathbf{H}\mathbf{x}^j - \mathbf{z}^j$, for example $\phi = \|\cdot\|_1$ will incur a fitting robust to outliers, $\phi = \|\cdot\|_2$ a fitting robust to noise; see [4, §6.1], hence our formulations are in some sense more robust than the ones in [10], [3].

VI. Approximate bilevel programs

Imposing $r_{\text{PD}} = r_{\text{KKT}} = 0$ in our formulations forces $\mathbf{x}^j$ to be a solution of MP($\mathcal{K}^j, F$), which incurs a *bilevel program*:

$$\begin{array}{ll} \min\limits_{F,\mathbf{x}} & \sum_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j) \\ \text{s.t.} & \mathbf{x}^j \text{ is a solution to MP}(\mathcal{K}^j, F) \quad \forall\, j \end{array} \tag{15}$$

This formulation has been proposed in [8] for inverse CO and in [16] for inverse VI. However, having the primal-dual or KKT systems as constraints is not practical because of the complementary constraints $F(\mathbf{x}|\boldsymbol{\theta})^T\mathbf{x} = \mathbf{b}^T\mathbf{y}$ or $\mathbf{x} \succeq 0$, $\boldsymbol{\pi} \succeq 0$, $\boldsymbol{\pi}^T\mathbf{x} = 0$, which cause the standard *Mangasarian-Fromovitz Constraint Qualification* (MFCQ) to be violated at any feasible point [18], hence generating severe numerical difficulties in standard nonlinear solvers [9], [11].

With $\mathcal{F} = \{F(\cdot, \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta}$ a parametric family, replacing $\phi(\mathbf{H}\mathbf{x} - \mathbf{z})$ by a general objective g, and having $\mathbf{A}, \mathbf{b}$ parametrized by $\boldsymbol{\theta}$, our formulations become:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\theta},\mathbf{x},\mathbf{y}} & [r_{\text{PD}}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}),\ g(\mathbf{x}, \boldsymbol{\theta})]^T \\ \text{s.t.} & \mathbf{A}(\boldsymbol{\theta})\mathbf{x} = \mathbf{b}(\boldsymbol{\theta}),\ \mathbf{A}(\boldsymbol{\theta})^T\mathbf{y} \preceq F(\mathbf{x}, \boldsymbol{\theta}) \\ & \boldsymbol{\theta} \in \Theta \end{array} \tag{16}$$

$$\begin{array}{ll} \min\limits_{\boldsymbol{\theta},\mathbf{x},\mathbf{y},\boldsymbol{\pi}} & [r_{\text{KKT}}(\mathbf{x}, \mathbf{y}, \boldsymbol{\pi}, \boldsymbol{\theta}),\ g(\mathbf{x}, \boldsymbol{\theta})]^T \\ \text{s.t.} & \mathbf{A}(\boldsymbol{\theta})\mathbf{x} = \mathbf{b}(\boldsymbol{\theta}),\ \mathbf{A}(\boldsymbol{\theta})^T\mathbf{y} \preceq F(\mathbf{x}, \boldsymbol{\theta}) \\ & \boldsymbol{\pi} \succeq 0,\ \boldsymbol{\theta} \in \Theta \end{array} \tag{17}$$

where Θ collects the admissible parameters and the residuals now depend on $\boldsymbol{\theta}$ via $F(\cdot, \boldsymbol{\theta})$. These programs are approximations of the bilevel program:

$$\begin{array}{ll} \min\limits_{\mathbf{p},\mathbf{x}} & g(\mathbf{x}, \boldsymbol{\theta}) \\ \text{s.t.} & \mathbf{x} \text{ is a solution to MP}(\mathcal{K}(\boldsymbol{\theta}), F(\cdot, \boldsymbol{\theta})) \end{array} \tag{18}$$

hence our formulation can be seen as a penalized reformulation of a bilevel program, in which we avoid to deal with the complementary condition. We note that bilevel programs are an important class of problems with many applications; see [11], [5]. Standard smoothing methods applied to $\mathbf{x} \succeq 0$, $\boldsymbol{\pi} \succeq 0$, $\mathbf{x}^T\boldsymbol{\pi} = 0$ are via the perturbed Fischer-Burmeister function [5, §6.5], but our smoothing via residual has a suboptimality interpretation, hence we call the novel problems (13) and (14) 'inverse VI and CO problems in *combined bilevel-residual* form', which are part or a larger and novel class of programs of the form (16) and (17) we call *approximate bilevel programs*.

VII. Residual and merit functions

In this section, we specify the residuals $r_{\text{PD}}, r_{\text{KKT}}$ and define more intuitive residuals associated to *suboptimal* solutions to VI($\mathcal{K}, F$), OP($\mathcal{K}, f$).

Definition 2. *Suppose $\mathcal{K}$ polyhedral given by (6). Given a map $F : \mathcal{K} \to \mathbb{R}^n$ and a function $f : \mathcal{K} \to \mathbb{R}$, the gap function associated to VI($\mathcal{K}, F$), the suboptimality gap*

associated to OP$(\mathcal{K}, f)$, and the duality gap associated to the primal-dual system (7) are defined by:

$$r_{VI}(\mathbf{x}) = \max_{\mathbf{u}\in\mathcal{K}} F(\mathbf{x})^T(\mathbf{x}-\mathbf{u}) \tag{19}$$

$$r_{OP}(\mathbf{x}) = f(\mathbf{x}) - \min_{\mathbf{u}\in\mathcal{K}} f(\mathbf{u}) \tag{20}$$

$$r_{PD}(\mathbf{x},\mathbf{y}) = F(\mathbf{x})^T\mathbf{x} - \mathbf{b}^T\mathbf{y} \tag{21}$$

These are classic bounds, analyzed respectively, *e.g.*, in [7, §3.1.5], [4, §9.3.1], and [3]. When primal feasibility holds, r_{OP} and r_{VI} are nonnegative and define *merit functions* [7, §1.5.3] because $r_{\text{VI}}(\mathbf{x}) = 0$ (resp. $r_{\text{OP}}(\mathbf{x}) = 0$) if and only if $\mathbf{x}$ is solution to VI$(\mathcal{K}, F)$ (resp. OP$(\mathcal{K}, f)$). Hence, we say that $\mathbf{x} \in \mathcal{K}$ is an *ϵ-suboptimal solution* to VI$(\mathcal{K}, F)$ (resp. OP$(\mathcal{K}, f)$) if $r_{\text{VI}}(\mathbf{x}) \leq \epsilon$ (resp. $r_{\text{OP}}(\mathbf{x}) \leq \epsilon$). In forward problems, r_{VI} and r_{OP} describe decisions that are at most ϵ from optimal. In reverse problems, they describe models that disagree with the data at most ϵ from the perfect fit.

While r_{VI} and r_{OP} are more intuitive, evaluating them requires solving a mathematical program, so they cannot be used in our MO formulation. Instead, r_{PD} is used. It is nonnegative when primal-dual feasibility holds, from weak LP duality; see [1], and so a merit function. We will show in the next section that r_{PD} defines the same bound as r_{VI}.

We now define the residual functions associated to the KKT system. We first define the slack variables associated to the dual feasibility condition:

$$\boldsymbol{\nu} := F(\mathbf{x}) - \mathbf{A}^T\mathbf{y} \tag{22}$$

which implies that dual feasibility is equivalent to $\boldsymbol{\nu} \succeq 0$.

Definition 3. *Suppose $\mathcal{K}$ is given by (6). Given a map $F : \mathcal{K} \to \mathbb{R}^n$, the residuals associated to the KKT system of VI$(\mathcal{K}, F)$ are given by (with $\circ$ is the Hadamard product):*

$$r_{stat}(\mathbf{x},\mathbf{y},\boldsymbol{\pi}) = F(\mathbf{x}) - \mathbf{A}^T\mathbf{y} - \boldsymbol{\pi} = \boldsymbol{\nu} - \boldsymbol{\pi} \tag{23}$$

$$r_{comp}(\mathbf{y},\boldsymbol{\pi}) = \mathbf{x} \circ \boldsymbol{\pi} = (x_i\pi_i)_{i=1}^n \tag{24}$$

Note that if $F = \nabla f$ for a potential f, then $r_{\text{stat}}, r_{\text{comp}}$ are the residuals associated to the KKT system of OP$(\mathcal{K}, f)$. Any nonnegative functions of $r_{\text{stat}}, r_{\text{comp}}$ that vanishes if and only if $r_{\text{stat}} = 0$, $r_{\text{comp}} = 0$ is a merit function. Common choices include ℓ_p norms, Hüber loss, log-barrier functions, which have different effects on the distribution of $r_{\text{stat}}, r_{\text{comp}}$ [4, §6.1]. We focus here on the KKT merit function when it is a ℓ_p norm of $r_{\text{stat}}, r_{\text{comp}}$ with a weighting factor $\alpha > 0$:

$$\begin{aligned} r_{\text{KKT}}^{\ell_p}(\mathbf{x},\mathbf{y},\boldsymbol{\pi}) &:= \|[\alpha\, r_{\text{stat}}(\mathbf{x},\mathbf{y},\boldsymbol{\pi}),\, r_{\text{comp}}(\mathbf{y},\boldsymbol{\pi})]\|_p \\ &= \left(\textstyle\sum_i \alpha^p|\nu_i - \pi_i|^p + |x_i\pi_i|^p\right)^{1/p} \end{aligned} \tag{25}$$

The above residual is used in practice as a certificate of suboptimality in iterative methods for solving OP$(\mathcal{K}, f)$ [6, §6]. It has been proposed in [10] as a measure of the fit of a CO model to the data.

VIII. Bounds on approximate solutions

We now derive necessary and sufficient conditions for ϵ-suboptimality. The results below, due to [1], compare r_{VI}, r_{PD}, and r_{OP}:

Theorem 4. *Suppose $\mathcal{K}$ given by (6). Let $\epsilon \geq 0$, $\mathbf{x} \in \mathcal{K}$. Then*

$$r_{VI}(\mathbf{x}) \leq \epsilon \iff \exists\, \mathbf{y} \,:\, \mathbf{A}^T\mathbf{y} \preceq F(\mathbf{x}),\, r_{PD}(\mathbf{x},\mathbf{y}) \leq \epsilon \tag{26}$$

In addition, if F is the gradient of a convex potential f:

$$r_{PD}(\mathbf{x}) \leq \epsilon \implies r_{OP}(\mathbf{x}) \leq \epsilon \tag{27}$$

Hence, when primal-duality feasibility holds, $r_{\text{PD}} \leq \epsilon$ is necessary and sufficient for ϵ-suboptimality for VI$(\mathcal{K}, F)$. When $f = \nabla F$, $r_{\text{PD}} \leq \epsilon$ is sufficient for ϵ-suboptimality for OP$(\mathcal{K}, f)$, but not necessary (see Appendix), hence r_{OP} defines weaker bounds than r_{VI} or r_{PD}. We now compare r_{VI} and r_{KKT}.

Theorem 5. *Suppose $\mathcal{K} \subseteq \mathbb{R}^n$ is given by (6). Let $\epsilon > 0$ and $\mathbf{x} \in \mathcal{K}$ such that $r_{VI}(\mathbf{x}) \leq \epsilon$. Then for all $p \geq 1$ and $\alpha > 0$:*

$$\exists\, \mathbf{y},\, \boldsymbol{\pi} \succeq 0 \quad : \quad \mathbf{A}^T\mathbf{y} \preceq F(\mathbf{x}),\, r_{KKT}^{\ell_p}(\mathbf{x},\mathbf{y},\boldsymbol{\pi}) \leq \epsilon \tag{28}$$

Reciprocally, if (28) holds for $p > 1$, then:

$$r_{VI}(\mathbf{x})/\epsilon \leq \left(1 + (\|\mathbf{x}\|_\infty/\alpha)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} n^{1-\frac{1}{p}} \tag{29}$$

and if (28) holds for $p = 1$, then:

$$r_{VI}(\mathbf{x})/\epsilon \leq (1 + (\|\mathbf{x}\|_\infty/\alpha - 1)_+) \tag{30}$$

We observe: 1) the upper bounds (29) and (30) are tight, 2) r_{VI} and r_{PD} define sufficient bounds for ϵ-suboptimality in the sense of r_{KKT}, but not necessary (see Appendix). In other words, r_{VI} and r_{PD} define stronger bounds than r_{KKT}. When $F = \nabla f$, combining Theorem 5 and $r_{\text{VI}} \geq r_{\text{OP}}$ from Theorem 4 gives bounds on the ratio of suboptimality of OP to ϵ-suboptimality under r_{KKT}:

Corollary 3. *Suppose $\mathcal{K}$ given by (6) and $F = \nabla f$ with f convex. Let $\epsilon > 0$, $\mathbf{x} \in \mathcal{K}$, $p > 1$ such that (28) holds, then:*

$$r_{OP}(\mathbf{x})/\epsilon \leq \left(1 + (\|\mathbf{x}\|_\infty/\alpha)^{\frac{p}{p-1}}\right)^{\frac{p-1}{p}} n^{1-\frac{1}{p}} \tag{31}$$

and if (28) holds for $p = 1$, then:

$$r_{OP}(\mathbf{x})/\epsilon \leq (1 + (\|\mathbf{x}\|_\infty/\alpha - 1)_+) \tag{32}$$

To summarize the results of this section, when primal-dual feasibility holds, we have, for any norm $\|\cdot\|$ on $\mathbb{R}^n$:

$$\begin{gathered} r_{\text{KKT}} \leq \epsilon \impliedby r_{\text{VI}} \leq \epsilon \iff r_{\text{PD}} \leq \epsilon \implies r_{\text{OP}} \leq \epsilon \\ r_{\text{OP}}/\epsilon = O(\|\mathbf{x}\|) \impliedby r_{\text{KKT}} \leq \epsilon \implies r_{\text{VI}}/\epsilon = O(\|\mathbf{x}\|) \end{gathered}$$

Consequently, r_{PD} given by (21) can be used to fit both VI and CO models to the data. If we solve the estimation problem and find that r_{PD} is small, then the fitting VI or CO model is consistent with the data. On the contrary, if many of the r_{PD} are large, then we can conclude that the fitting VI is not a good model for our data. However, this is not sufficient to say that our CO model is inconsistent with the data because large values of r_{PD} can be caused by the existence of optimal solutions with large norms. Values of r_{KKT} provide more insights on whether the CO model can explain the data.

In many scenarios, the feasible set $\mathcal{K}$ is bounded, *e.g.* in game theory where $\mathcal{K}$ describes strategy distributions [14],

hence r_{VI}, r_{OP}, r_{PD}, r_{KKT} are equivalent (in the sense of norms), and r_{PD} may be necessary and sufficient to measure the fit of VI and CO models to the data.

IX. Scalarization methods for estimation

We recall the pairs of objectives in (13) and (14), where the dependencies on F are made explicit:

$$[\,\Sigma_j r_{\text{PD}}(\mathbf{x}^j, \mathbf{y}^j, F),\ \Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j)\,]^T \tag{33}$$

$$[\,\Sigma_j r_{\text{KKT}}(\mathbf{x}^j, \mathbf{y}^j, \boldsymbol{\pi}^j, F),\ \Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j)\,]^T \tag{34}$$

In this section, we apply the common *weighted sum method* to articulate preferences between the two objectives in (33) (resp. (34)); see [12] for a survey on MO optimization. It is known that the weighted sum method is *sufficient* for Pareto optimality, but varying the weights $\mathbf{w}$ may not capture all the Pareto optimal points, although it provides information about available trade-offs between the objectives. For instance, with some tuning on $\mathbf{w}$, our formulations (13) and (14) capture, in some sense, the optimal solutions to the inverse problems (12) and (11) proposed in [10] and [3]. More precisely, if the objective in (12) or (11) attains values less than ϵ, we can achieve, for all $\epsilon' > 0$:

$$\Sigma_j r_{\text{PD}}(\mathbf{x}^j, \mathbf{y}^j, F) \le \epsilon, \quad \Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j) \le \epsilon' \tag{35}$$

$$\Sigma_j r_{\text{KKT}}(\mathbf{x}^j, \mathbf{y}^j, \boldsymbol{\pi}^j, F) \le \epsilon, \quad \Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j) \le \epsilon' \tag{36}$$

with the following scalarization of (13) and (14):

$$\begin{array}{ll} \min\limits_{F,\mathbf{x},\mathbf{y}} & w_{\text{MP}}\, \Sigma_j r_{\text{PD}} + w_{\text{obs}}\, \Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j) \\ \text{s.t.} & \mathbf{A}^j\mathbf{x}^j = \mathbf{b}^j,\ \mathbf{x}^j \succeq 0 \quad \forall\, j \\ & \mathbf{A}^{jT}\mathbf{y}^j \preceq F(\mathbf{x}^j) \qquad \forall\, j \\ & F \in \mathcal{F} \end{array} \tag{37}$$

$$\begin{array}{ll} \min\limits_{F,\mathbf{x},\mathbf{y},\boldsymbol{\pi}} & w_{\text{MP}}\, \Sigma_j r_{\text{KKT}} + w_{\text{obs}}\, \Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j) \\ \text{s.t.} & \mathbf{A}^j\mathbf{x}^j = \mathbf{b}^j,\ \mathbf{x}^j \succeq 0 \qquad \forall\, j \\ & \mathbf{A}^{jT}\mathbf{y}^j \preceq F(\mathbf{x}^j),\ \boldsymbol{\pi}^j \succeq 0 \quad \forall\, j \\ & F \in \mathcal{F} \end{array} \tag{38}$$

Theorem 6. *Let $\epsilon \ge 0$ and suppose the minimum objective value of the inverse problem (11) (resp. (12)) is ϵ and attained at a primal feasible point. Then, for all $\epsilon' > 0$ and for all $w_{MP}, w_{obs} > 0$ with $w_{MP} + w_{obs} = 1$ and $w_{obs} \ge \frac{\epsilon}{\epsilon+\epsilon'}$, problem (37) (resp. (38)) is sufficient for (35) (resp. (36)).*

Observe that setting $\epsilon = 0$ in Theorem 6 gives the following result, which characterizes a perfect fit to the data:

Corollary 4. *Suppose the optimal objective value of the inverse problem (11) (resp. (12)) is null. Then for all weights $w_{MP}, w_{obs} > 0$ with $w_{MP}+w_{obs} = 1$, problem (37) (resp. (38)) is sufficient for (35) (resp. (36)) with $\epsilon = \epsilon' = 0$.*

The result in Theorem 6 suggests to set w_{obs} close to 1 if we trust our observations. It requires that an optimal solution to (12) or (11) is primal-feasible. For example, the absence of noise in the constraints $\mathbf{A}^j\mathbf{x} = \mathbf{b}^j$ and in the observations $\mathbf{z}^j = \mathbf{H}\mathbf{x}^j$ is sufficient to have this condition.

X. Scalarization methods for control

Similarly, if there exists a pair of control policy and induced response $(\boldsymbol{\theta}, \mathbf{x})$ such that the upper-level objective in (18) is equal to α, then, for all $\epsilon' > 0$, we can achieve:

$$r_{\text{PD}}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) \le \epsilon', \quad g(\mathbf{x}, \boldsymbol{\theta}) \le \alpha \tag{39}$$

$$r_{\text{KKT}}(\mathbf{x}, \mathbf{y}, \boldsymbol{\pi}, \boldsymbol{\theta}) \le \epsilon', \quad g(\mathbf{x}, \boldsymbol{\theta}) \le \alpha \tag{40}$$

with the following scalarization of the approximate bilevel programs (16) and (17):

$$\begin{array}{ll} \min\limits_{\boldsymbol{\theta},\mathbf{x},\mathbf{y}} & w_{\text{MP}}\, r_{\text{PD}}(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) + w_{\text{g}}\, g(\mathbf{x}, \boldsymbol{\theta}) \\ \text{s.t.} & \mathbf{A}(\boldsymbol{\theta})\mathbf{x} = \mathbf{b}(\boldsymbol{\theta}),\ \mathbf{A}(\boldsymbol{\theta})^T\mathbf{y} \preceq F(\mathbf{x}, \boldsymbol{\theta}) \end{array} \tag{41}$$

$$\begin{array}{ll} \min\limits_{\boldsymbol{\theta},\mathbf{x},\mathbf{y},\boldsymbol{\pi}} & w_{\text{MP}}\, r_{\text{KKT}}(\mathbf{x}, \mathbf{y}, \boldsymbol{\pi}, \boldsymbol{\theta}) + w_{\text{g}}\, g(\mathbf{x}, \boldsymbol{\theta}) \\ \text{s.t.} & \mathbf{A}(\boldsymbol{\theta})\mathbf{x} = \mathbf{b}(\boldsymbol{\theta}),\ \mathbf{A}(\boldsymbol{\theta})^T\mathbf{y} \preceq F(\mathbf{x}, \boldsymbol{\theta}) \\ & \boldsymbol{\pi} \succeq 0 \end{array} \tag{42}$$

Theorem 7. *Let $\alpha > 0$ and MP$(\mathcal{K}, F)$ refers to VI$(\mathcal{K}, F)$. Suppose the upper-level objective g in (18) is a nonnegative function and the minimum objective value of (18) is α (it may not be attained). Then, for all $\epsilon' > 0$ and for all weights $w_{MP}, w_{obs} > 0$ with $w_{MP} + w_{obs} = 1$ and $w_{MP} \ge \frac{\alpha}{\alpha+\epsilon'}$, the problem (41) (resp. (42)) is sufficient for (39) (resp. (40)).*

Hence our approximate bilevel programs are a novel unifying framework for solving bilevel programs and inverse VI/CO problems. With enough weight on the observation residuals, they can be used to impute the function that describes a VI or CO model, and with enough weight on r_{PD} or r_{KKT}, they can be used to solve bilevel programs.

Finally, it is desirable to divide the objective functions by their maximum value (approximated via engineering knowledge) to have a consistent comparison between them.

XI. Numerical results

Traffic assignment: We consider the same highway network and experimental setup as in [16]. Recall the aggregate flow $\mathbf{v} = (v_a)_{a\in\mathcal{A}} = \mathbf{Z}\mathbf{x} = \sum_k \mathbf{x}^k$ with $\mathbf{x}^k$ the commodity flows. Our parametric VI model $F(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{Z}^T S(\mathbf{Z}\mathbf{x}, \boldsymbol{\theta})$ has polynomial delays $S(\cdot, \boldsymbol{\theta}) : \mathbb{R}^{\mathcal{A}}_+ \to \mathbb{R}$ is:

$$S(\mathbf{v}, \boldsymbol{\theta}) = \Big(d_a(1 + \textstyle\sum_{i=1}^6 \theta_i (v_a/m_a)^i)\Big)_{a\in\mathcal{A}} \tag{43}$$

for all $\boldsymbol{\theta} \in \Theta := \mathbb{R}^6_+$, where d_a and m_a are the free flow delay and capacity on arc a. We note that $F(\cdot, \boldsymbol{\theta})$ is nonnegative, monotone, convex for all $\boldsymbol{\theta} \in \Theta$. Using our inverse VI formulation (37) with $\phi = \frac{1}{2}\|\cdot\|_2^2$, we want to impute $\boldsymbol{\theta}$ from $N = 4$ partial observations $\mathbf{z}^j = \tilde{\mathbf{H}}\mathbf{v}^j \in \mathbb{R}^{\mathcal{A}^{\text{obs}}}_+$ of UE aggregate flows $\mathbf{v}^j$ associated to four demand vectors $\mathbf{b}^j \in \mathbb{R}^{|\mathcal{C}||\mathcal{N}|}$, $j = 1, 2, 3, 4$. The measurements are obtained by solving the traffic assignment problem (1) with two 'true' delay functions: (44) estimated by the Bureau of Public Roads and the hyperbolic delay (45):

$$S^{\text{poly}}(\mathbf{v}) = \big(d_a(1 + 0.15(v_a/m_a)^4)\big)_{a\in\mathcal{A}} \tag{44}$$

$$S^{\text{hyper}}(\mathbf{v}) = (1 - 3.5/3 + 3.5/(3 - v_a/m_a))_{a\in\mathcal{A}} \tag{45}$$

We scale $\Sigma_j r_{\text{PD}}(\mathbf{x}^j, \mathbf{y}^j, \boldsymbol{\theta})$ and $\Sigma_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j)$ by the inverse of $\max \sum_j F(\mathbf{x}^j)^T\mathbf{x}^j \approx \sum_j \sum_{k\in\mathcal{C}} 5\,\lambda_k c_k$ and

$\max \sum_j \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j) \approx \sum_j \phi(\mathbf{H}\mathbf{x}_0^j - \mathbf{z}^j)$ where $\mathbf{x}_0^j$ is the UE flow solution to VI($\mathcal{K}^j, F(\cdot, \boldsymbol{\theta}_0)$) with initial parameters $\boldsymbol{\theta}_0$, λ_k is the demand rate in commodity $(s_k, t_k) \in \mathcal{C}$, and c^k is the shortest path cost between s_k and t_k with free flow delays d_a.[1] The program (37) is solved via BCD where each block is updated using CVXOPT.[2] Figure 2 provides the final values of the *scaled* duality gap $\Sigma_j r_{\text{PD}}(\mathbf{x}^j, \mathbf{y}^j, \boldsymbol{\theta})$ and observation residual $\Sigma_i \phi(\mathbf{H}\mathbf{x}^j - \mathbf{z}^j)$ for different values of w_{obs} and $w_{\text{MP}} = 1 - w_{\text{obs}}$.

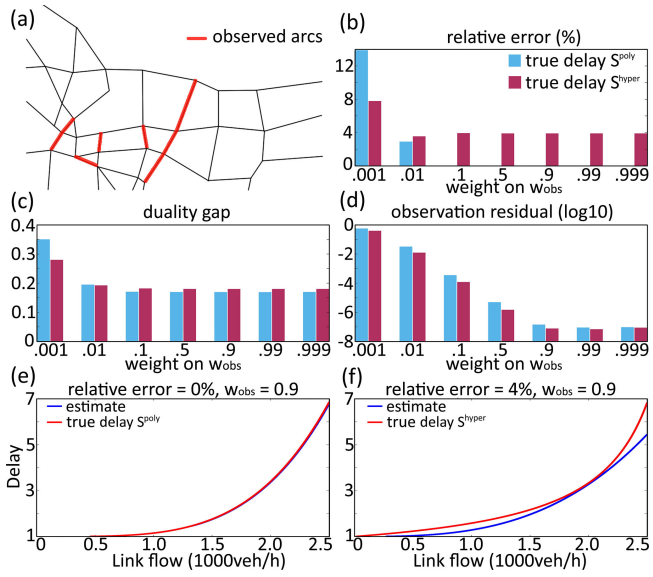

Fig. 1. **Imputation of the delay maps S^{poly}, S^{hyper} using formulation (37) with parametric model map $F(\cdot, \boldsymbol{\theta}) = \mathbf{Z}^T S(\mathbf{Z}\mathbf{x}, \boldsymbol{\theta})$ given by (43). As predicted by Theorem 6, the observation residual decreases (to 10^{-7}) as w_{obs} gets close to 1 as shown in (d), while the duality gap stays constant as shown in (c). The relative error on the flow predicted by the imputed map $F(\cdot, \hat{\boldsymbol{\theta}})$ is small for w_{obs} large enough as shown in (b). With accurate measurements, we suggest to solve (37) with $w_{\text{obs}} = 0.9$, which gives the estimated $1 + \sum_{i=1}^6 \hat{\theta}_i u^i$ in (e) and (f).**

Consumer utility: As in [10], we consider $n = 5$ firms sharing the market. We are firm 3 and we observe $N = 200$ pairs $(\mathbf{p}^j, \mathbf{z}^j)$, $j = 1, \cdots, N$ with $\mathbf{z}^j = [x_2^j, \cdots, x_5^j]^T$ where $\mathbf{x}^j \in \mathbb{R}_+^5$ is the consumer demand incurred by prices $\mathbf{p}^j$ using the CO model (3) with 'real' consumer utility U^{real} given by (46). The prices are sampled uniformly in $[8, 12]^5$. We impute U^{real} using the parametric utility given by (47) with $(\mathbf{Q}, \mathbf{r}) \in \Theta := \{(\mathbf{Q}, \mathbf{r}) \,|\, Q\mathbf{x}_{\max} + \mathbf{r} \succeq 0,\ \mathbf{r} \succeq 0,\ \mathbf{Q} \preceq 0\}$ so that U is concave quadratic and nondecreasing on the demand range $[0, \mathbf{x}_{\max}]$, where $\mathbf{x}_{\max}$ is obtained from the data. The imputed utility is used by firm 3 to price its product to achieve different demand levels, given other prices sampled uniformly in $[8, 12]$. The numerical results are shown in Figure 4 with 2 models for $\mathbf{A} = 50(\mathbf{I} + \mathbf{B})$: *model 1* where $\mathbf{B}_{ij}$ is sampled uniformly in $[0, 0.3]$ for $i \neq j$, and *model 2* where $\mathbf{B}_{ij}$ is sampled from $0.5 \cdot$Bernoulli(0.3).

[1]We assume the maximum delay is at most 5 times the free flow delay.

[2]CVXOPT is a Python software package available at `http://cvxopt.org`. Implementation of the block descent is open source and available at: `https://github.com/jeromethai/traffic-estimation-wardrop`.

$$U^{\text{real}}(\mathbf{x}) = \mathbf{1}^T \sqrt{\mathbf{A}\mathbf{x} + \mathbf{b}} \tag{46}$$

$$U(\mathbf{x}, \mathbf{Q}, \mathbf{r}) = (1/2)\mathbf{x}^T \mathbf{Q}\mathbf{x} + \mathbf{r}^T \mathbf{x} \tag{47}$$

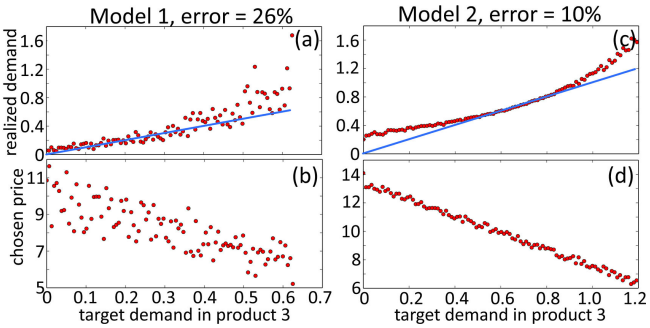

Fig. 2. **Use of the imputed utility to price product 3 for different target demands x_3^{des}. In (b), the prices are scattered due to correlations with other prices in model 1, while in (d), the prices vary linearly with x_3^{des} since the prices in model 2 are more uncorrelated. In (a), (c) the blue line is the $x = y$ line. For both models, the imputed utility performs well with relative errors of 26% and 10% on the training data and target demands $\mathbf{x}_3^{\text{des}}$ close to realized demands $\mathbf{x}_3^{\text{real}}$.**

References

[1] M. Aghassi, D. Bertsimas, and G. Perakis. Solving asymmetric variational inequalities via convex optimization. *Operations Research Letters 34*, 5:481–490, 2006.

[2] M. Beckmann, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*. Cowles Commission Monograph, 1956.

[3] D. Bertsimas, V. Gupta, and I. Ch. Paschalidis. Data-Driven Estimation in Equilibrium Using Inverse Optimization. *Mathematical Programming*, 2014.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 8 2004.

[5] S. Dempe. *Foundations of Bilevel Programming*. Springer, 2002.

[6] A. S. El-Bakry, R. A. Tapia, T. Tsuchiya, and Y. Zhang. On the Formulation and Theory of the Newton Interior-Point Method for Nonlinear Programming. *Journal of Optimization Theory and Applications*, 89:507–541, 1996.

[7] F. Facchinei and J. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, New York, 2003.

[8] G. Iyengar and W. Kang. Inverse conic programming with applications. *Operations Research Letters 33*, 2005.

[9] H. Jiang, D. Ralph, and J. Pang. QPECgen, a MATLAB generator for mathematical programs with quadratic objectives and affine variational inequality constraints. *Computational Optimization and Applications*, 13:25–59, 1999.

[10] A. Keshavarz, Y. Wang, and S. Boyd. Imputing a Convex Objective Function. *IEEE International Symposium on Intelligent Control (ISIC)*, 2011.

[11] Z. Q. Luo, J. S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. Cam, 1996.

[12] R. T. Marler and J. S. Arora. Survey of multi-objective optimization methods for engineering. *Structural and Multidisciplinary Optimization*, pages 369–395, 2004.

[13] M. Patriksson. *The Traffic Assignment Problem - Models and Methods*. VSP, Utrecht, 1994.

[14] W. H. Sandholm. Potential Games with Continuous Player Sets. *Journal of Economic Theory*, 97:81–108, 2001.

[15] Gesualdo Scutari, Daniel P. Palomar, Francisco Facchinei, and Jong-Shi Pang. Convex Optimization, Game Theory, and Variational Inequality Theory. *IEEE Signal Processing Magazine*, 35, 2010.

[16] J. Thai, R. Hariss, and A. Bayen. A Multi-Convex approach to Latency Inference and Control in Traffic Equilibria from Sparse data. *Submitted to the 2015 American Control Conference*, 2014.

[17] J. G. Wardrop and J. I. Whitehead. Correspondence. Some Theoretical Aspects of Road Traffic Research. *ICE Proceedings: Engineering Divisions 1*, 1952.

[18] J. J. Ye and D. L. Zhu. Optimality conditions for bilevel programming problems. *Optimization: A Journal of Mathematical Programming and Operations Research*, 33, 1995.