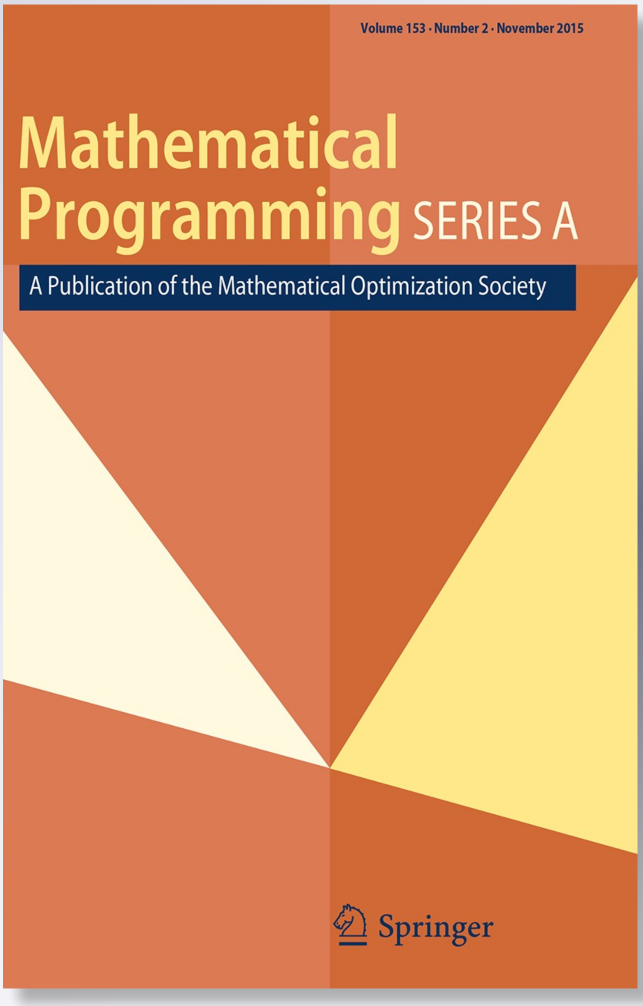
Data-driven estimation in equilibrium using inverse optimization

Dimitris Bertsimas, Vishal Gupta & Ioannis Ch. Paschalidis

Mathematical Programming
A Publication of the Mathematical Optimization Society

ISSN 0025-5610
Volume 153
Number 2

Math. Program. (2015) 153:595-633
DOI 10.1007/s10107-014-0819-4


Volume 153 · Number 2 · November 2015
Mathematical Programming SERIES A
A Publication of the Mathematical Optimization Society
Springer

Springer

Math. Program., Ser. A (2015) 153:595–633
DOI 10.1007/s10107-014-0819-4

FULL LENGTH PAPER

Data-driven estimation in equilibrium using inverse optimization

Dimitris Bertsimas · **Vishal Gupta** ·
Ioannis Ch. Paschalidis

Received: 23 November 2012 / Accepted: 12 September 2014 / Published online: 30 September 2014

Abstract Equilibrium modeling is common in a variety of fields such as game theory and transportation science. The inputs for these models, however, are often difficult to estimate, while their outputs, i.e., the equilibria they are meant to describe, are often directly observable. By combining ideas from inverse optimization with the theory of variational inequalities, we develop an efficient, data-driven technique for estimating the parameters of these models from observed equilibria. We use this technique to estimate the utility functions of players in a game from their observed actions and to estimate the congestion function on a road network from traffic count data. A distinguishing feature of our approach is that it supports both parametric and *nonparametric* estimation by leveraging ideas from statistical learning (kernel methods and regularization operators). In computational experiments involving Nash and Wardrop equilibria in a nonparametric setting, we find that a) we effectively estimate the unknown demand or congestion function, respectively, and b) our proposed regularization technique substantially improves the out-of-sample performance of our estimators.

D. Bertsimas (✉)
MIT, Sloan School of Management, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA
e-mail: dbertsim@mit.edu

V. Gupta
Operations Research Center, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA
e-mail: vgupta1@mit.edu

I. Ch. Paschalidis
Department of Electrical and Computer Engineering, Boston University,
Boston, MA 02215, USA
e-mail: yannisp@bu.edu

Keywords Equilibrium · Nonparametric estimation · Utility estimation · Traffic assignment

Mathematics Subject Classification 74G75 Equilibrium: Inverse Problems · 62G05 Nonparametric Inference: Estimation · 62P20 Applications to Economics · 90B20 Operations Research and Management Science: Traffic Problems

1 Introduction

Modeling phenomena as equilibria is a common approach in a variety of fields. Examples include Nash equilibrium in game theory, traffic equilibrium in transportation science and market equilibrium in economics. Often, however, the model primitives or "inputs" needed to calculate equilibria are not directly observable and can be difficult to estimate. Small errors in these estimates may have large impacts on the resulting equilibrium. This problem is particularly serious in *design* applications, where one seeks to (re)design a system so that the induced equilibrium satisfies some desirable properties, such as maximizing social welfare. In this case, small errors in the estimates may substantially affect the optimal design. Thus, developing accurate estimates of the primitives is crucial.

In this work we propose a novel framework to estimate the unobservable model primitives for systems in equilibrium. Our data-driven approach hinges on the fact that although the model primitives may be unobservable, it is frequently possible to observe equilibria experimentally. We use these observed equilibria to estimate the original primitives.

We draw on an example from game theory to illustrate. Typically, one specifies the utility functions for each player in a game and then calculates Nash equilibria. In practice, however, it is essentially impossible to observe utilities directly. Worse, the specific choice of utility function often makes a substantial difference in the resulting equilibrium. Our approach amounts to estimating a player's utility function from her actions in previous games, assuming her actions were approximately equilibria with respect to her opponents. In contrast to her utility function, her previous actions *are* directly observable. This utility function can be used either to predict her actions in future games, or as an input to subsequent mechanism design problems involving this player in the future.

A second example comes from transportation science. Given a particular road network, one typically specifies a cost function and then calculates the resulting flow under user (Wardrop) equilibrium. However, measuring the cost function directly in a large-scale network is challenging because of the interdependencies among arcs. Furthermore, errors in estimates of cost functions can have severe and counterintuitive effects; Braess paradox (see [13]) is one well-known example. Our approach amounts to estimating cost functions using current traffic count data (flows) on the network, assuming those flows are approximately in equilibrium. Again, in contrast to the cost function, traffic count data are readily observable and frequently collected on many real-life networks. Finally, our estimate can be used either to predict congestion on the network in the future, or else to inform subsequent network design problems.

In general, we focus on equilibria that can be modeled as the solution to a variational inequality (VI). VIs are a natural tool for describing equilibria with examples spanning economics, transportation science, physics, differential equations, and optimization. (See Sect. 2.1 or [26] for detailed examples.) Our model centers on solving an *inverse variational inequality problem*: given data that we believe are equilibria, i.e., solutions to some VI, estimate the function which describes this VI, i.e., the model primitives.

Our formulation and analysis is motivated in many ways by the inverse optimization literature. In inverse optimization, one is given a candidate solution to an optimization problem and seeks to characterize the cost function or other problem data that would make that solution (approximately) optimal. See [27] for a survey of inverse combinatorial optimization problems, [3] for the case of linear optimization and [28] for the case of conic optimization. The critical difference, however, is that we seek a cost function that would make the observed data equilibria, not optimal solutions to an optimization problem. In general, optimization problems can be reformulated as variational inequalities (see Sect. 2.1), so that our inverse VI problem *generalizes* inverse optimization, but this generalization allows us to address a variety of new applications.

To the best of our knowledge, we are the first to consider inverse variational inequality problems. Previous work, however, has examined the problem of estimating parameters for systems assumed to be in equilibrium, most notably the structural estimation literature in econometrics and operations management ([4,5,32,35]). Although there are a myriad of techniques collectively referred to as structural estimation, roughly speaking, they entail (1) assuming a parametric model for the system including probabilistic assumptions on random quantities, (2) deducing a set of necessary (structural) equations for unknown parameters, and, finally, (3) solving a constrained optimization problem corresponding to a generalized method of moments (GMM) estimate for the parameters. The constraints of this optimization problem include the structural equations and possibly other application-specific constraints, e.g., orthogonality conditions of instrumental variables. Moreover, this optimization problem is typically difficult to solve numerically, as it can be non-convex with large flat regions and multiple local optima (see [4] for some discussion).

Our approach differs from structural estimation and other specialized approaches in a number of respects. From a philosophical point of view, the most critical difference is in the objective of the methodology. Specifically, in the structural estimation paradigm, one posits a "ground-truth" model of a system with a known parametric form. The objective of the method is to learn the parameters in order to provide insight into the system. By contrast, in our paradigm, we make no assumptions (parametric or nonparametric) about the true mechanics of the system; we treat is as a "black-box." Our objective is to fit a model—in fact, a VI—that can be used to predict the behavior of the system. We make no claim that this fitted model accurately reflects "reality," merely that it has good predictive power.

This distinction is subtle, mirroring the distinction between "data-modelling" in classical statistics and "algorithmic modeling" in machine learning. (A famous, albeit partisaned, account of this distinction is [15].) Our approach is kindred to the machine learning point of view. For a more detailed discussion, please see Appendix 2.

This philosophical difference has a number of *practical* consequences:

1. **Minimal probabilistic assumptions:** Our method has provably good performance in a very general setting with minimal assumptions on the underlying mechanism generating the data. (See Theorems 6–8 for precise statements.) By contrast, other statistical methods, including structural estimation, require a full-specification of the data generating mechanism and can yield spurious results if this specification is inaccurate.
2. **Tractability:** Since our fitted model need not correspond exactly to the underlying system dynamics, we have considerably more flexibility in choosing its functional form. For several interesting choices, including nonparametric specifications (see next point), the resulting inverse VI problem can be reformulated as a conic optimization problem. Conic optimization problems are both theoretically and numerically tractable, even for large scale instances ([12]), in sharp contrast to the non-convex problems that frequently arise in other methods.
3. **Nonparametric estimation:** Like existing methods in inverse optimization and structural estimation, our approach can be applied in a parametric setting. Unlike these approaches, our approach also extends naturally to a nonparametric description of the function $\mathbf{f}$ defining the VI. To the best of our knowledge, existing methods do not treat this possibility. Partial exceptions are [8] and [25] which use nonparametric estimators for probability densities, but parametric descriptions of the mechanism governing the system. The key to our nonparametric approach is to leverage kernel methods from statistical learning to reformulate the infinite dimensional inverse variational inequality problem as a finite dimensional, convex quadratic optimization problem. In applications where we may not know, or be willing to specify a particular form for $\mathbf{f}$ we consider this non-parametric approach particularly attractive.

Although there are other technical differences between these approaches—for example, some structural estimation techniques can handle discrete features while our method applies only to continuous problems—we feel that the most important difference is the aforementioned intended purpose of the methodology. We see our approach as complementary to existing structural estimation techniques and believe in some applications practitioners may prefer it for its computational tractability and relatively fewer modeling assumptions. Of course, in applications where the underlying assumptions of structural estimations or other statistical techniques are valid, those techniques may yield potentially stronger claims about the underlying system.

We summarize our contributions below:

1. We propose the inverse variational inequality problem to model inverse equilibrium. We illustrate the approach by estimating market demand functions under Bertrand-Nash equilibrium and by estimating the congestion function in a traffic equilibrium.
2. We formulate an optimization problem to solve a parametric version of the inverse variational inequality problem. The complexity of this optimization depends on the particular parametric form of the function to be estimated. We show that for several interesting choices of parametric form, the parametric version of the inverse variational inequality problem can be reformulated as a simple conic optimization problem.

3. We formulate and solve a nonparametric version of the inverse variational inequality problem using kernel methods. We show that this problem can be efficiently solved as a convex quadratic optimization problem whose size scales linearly with the number of observations.
4. Under very mild assumptions on the mechanism generating the data, we show that both our parametric and non-parametric formulations enjoy a strong generalization guarantee similar to the guarantee enjoyed by other methods in machine learning. Namely, if the fitted VI explains the existing data well, it will continue to explain new data well. Moreover, under some additional assumptions on the optimization problem, equilibria from the VI serve as good predictions for new data points.
5. We provide computational evidence in the previous two examples—demand estimation under Nash equilibrium and congestion function estimation under traffic equilibrium—that our proposed approach recovers reasonable functions with good generalization properties and predictive power. We believe these results may merit independent interest in the specialized literature for these two applications.

The remainder of this paper is organized as follows. Section 2 reviews background material on equilibrium modeling through VIs. Section 3 formally defines the inverse variational inequality problem and solves it in the case that the function to be estimated has a known parametric form. In preparation for the nonparametric case, Sect. 4 reviews some necessary background material on kernels. Section 5 formulates and solves the nonparametric inverse variational inequality problem using kernels, and Sect. 6 illustrates how to incorporate priors, semi-parametric modeling and ambiguity sets into this framework. Section 7 states our results on the generalization guarantees and predictive power of our approach. Finally, Sect. 8 presents some computational results, and Sect. 9 concludes. In the interest of space, almost all proofs are placed in the Appendix.

In what follows we will use boldfaced capital letters (e.g., $\mathbf{A}$, $\mathbf{W}$) to denote matrices, boldfaced lowercase letters (e.g., $\mathbf{x}$, $\mathbf{f}(\cdot)$) to denote vectors or vector-valued functions, and ordinary lowercase letters to denote scalars. We will use caligraphic capital letters (e.g., $\mathcal{S}$) to denote sets. For any proper cone C, i.e. C is pointed, closed, convex and has a strict interior, we will say $\mathbf{x} \leq_C \mathbf{y}$ whenever $\mathbf{y} - \mathbf{x} \in C$.

2 Variational inequalities: background

2.1 Definitions and examples

In this section, we briefly review some results on variational inequalities that we use in the remainder of the paper. For a more complete survey, see [26].

Given a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ and a non-empty set $\mathcal{F} \subseteq \mathbb{R}^n$ the variational inequality problem, denoted VI($\mathbf{f}, \mathcal{F}$), is to find an $\mathbf{x}^* \in \mathcal{F}$ such that

$$\mathbf{f}(\mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*) \geq 0, \quad \forall \mathbf{x} \in \mathcal{F}. \tag{1}$$

A solution $\mathbf{x}^*$ to VI($\mathbf{f}, \mathcal{F}$) need not exist, and when it exists, it need not be unique. We can guarantee the existence and uniqueness of the solution by making appropriate assumptions on $\mathbf{f}(\cdot)$ and/or $\mathcal{F}$, e.g., $\mathbf{f}$ continuous and $\mathcal{F}$ convex and compact. See [26] for other less stringent conditions.

There are at least three classical applications of VI modeling that we will refer to throughout the paper: constrained optimization, Nash Equilibrium, and Traffic (or Market) Equilibrium.

Constrained Optimization The simplest example of a VI is in fact not an equilibrium, per se, but rather convex optimization. Nonetheless, the specific example is very useful in building intuition about VIs. Moreover, using this formalism, one can derive many of the existing results in the inverse optimization literature as a special case of our results for inverse VIs in Sect. 3.2.

Consider the problem

$$\min_{\mathbf{x}\in\mathcal{F}} F(\mathbf{x}). \tag{2}$$

The first order necessary conditions for an optimal solution of this problem are (see, e.g., [11])

$$\nabla F(\mathbf{x}^*)^T(\mathbf{x}-\mathbf{x}^*) \geq 0, \quad \forall \mathbf{x}\in\mathcal{F}. \tag{3}$$

These conditions are sufficient in the case that F is a convex function, and $\mathcal{F}$ is a convex set. Observe, then, that solving (2) is equivalent to finding a point which satisfies Eq. (3), which is equivalent to solving VI(∇F, $\mathcal{F}$).

Note that, in general, a VI with a function $\mathbf{f}$ whose Jacobian is symmetric models an optimization problem (see [26]).

Nash Equilibrium Our first application of VI to model equilibrium is non-cooperative Nash equilibrium. Consider a game with p players. Each player i chooses an action from a set of feasible actions, $\mathbf{a}_i \in \mathcal{A}_i \subseteq \mathbb{R}^{m_i}$, and receives a utility $U_i(\mathbf{a}_1, \ldots, \mathbf{a}_p)$. Notice in particular, that player i's payoff may depend upon the actions of other players. We will assume that U_i is differentiable and concave in $\mathbf{a}_i$ for all i and that $\mathcal{A}_i$ is convex for all i.

A profile of actions for the players $(\mathbf{a}_1^*, \mathbf{a}_2^*, \ldots \mathbf{a}_p^*)$ is said to be a Nash Equilibrium if no single player can unilaterally change her action and increase her utility. See [22] for a more complete treatment. In other words, player i plays her best response given the actions of the other players. More formally,

$$\mathbf{a}_i^* \in \arg\max_{\mathbf{a}\in\mathcal{A}_i} U_i(\mathbf{a}_1^*, \ldots, \mathbf{a}_{i-1}^*, \mathbf{a}, \mathbf{a}_{i+1}^*, \ldots, \mathbf{a}_p^*), \quad i = 1, \ldots, p. \tag{4}$$

This condition can be expressed as a VI. Specifically, a profile $\mathbf{a}^* = (\mathbf{a}_1^*, \mathbf{a}_2^*, \ldots \mathbf{a}_p^*)$ is a Nash Equilibrium, if and only if it solves VI($\mathbf{f}$, $\mathcal{F}$) where $\mathcal{F} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_p$,

$$\mathbf{f}(\mathbf{a}) = \begin{pmatrix} -\nabla_1 U_1(\mathbf{a}) \\ \vdots \\ -\nabla_p U_p(\mathbf{a}) \end{pmatrix} \tag{5}$$

and ∇_i denotes the gradient with respect to the variables $\mathbf{a}_i$ (see [26] for a proof.)

It is worth pointing out that many authors use Eq. (4) to conclude

$$\nabla_i U_i(\mathbf{a}_1^*, \ldots, \mathbf{a}_p^*) = \mathbf{0}, \quad i = 1, \ldots, p, \tag{6}$$

where ∇_i refers to a gradient with respect to the coordinates of $\mathbf{a}_i$. This characterization assumes that each player's best response lies on the strict interior of her strategy set $\mathcal{A}_i$. The assumption is often valid, usually because the strategy sets are unconstrained. Indeed, this condition can be derived as a special case of (5) in the case $\mathcal{A}_i = \mathbb{R}^{m_i}$. In some games, however, it is not clear that an equilibrium must occur in the interior, and we must use (5) instead. We will see an example in Sect. 3.2.

Wardrop Equilibrium Our final example of a VI is Wardrop or user-equilibrium from transportation science. Wardrop equilibrium is extremely close in spirit to the market (Walrasian) equilibrium model in economics—see [19,41]—and our comments below naturally extend to the Walrasian case.

Specifically, we are given a directed network of nodes and arcs $(\mathcal{V}, \mathcal{A})$, representing the road network of some city. Let $\mathbf{N} \in \{0, 1\}^{|\mathcal{V}| \times |\mathcal{A}|}$ be the node-arc incidence matrix of this network. For certain pairs of nodes $\mathbf{w} = (w_s, w_t) \in \mathcal{W}$, we are also given an amount of flow $d^{\mathbf{w}}$ that must flow from w_s to w_t. The pair $\mathbf{w}$ is referred to as an origin–destination pair. Let $\mathbf{d}^{\mathbf{w}} \in \mathbb{R}^{|V|}$ be the vector which is all zeros, except for a $(-d^{\mathbf{w}})$ in the coordinate corresponding to node w_s and a $(d^{\mathbf{w}})$ in the coordinate corresponding to node w_t.

We will say that a vector of flows $\mathbf{x} \in \mathbb{R}_+^{|\mathcal{A}|}$ is feasible if $\mathbf{x} \in \mathcal{F}$ where

$$\mathcal{F} = \left\{ \mathbf{x} : \exists \mathbf{x}^{\mathbf{w}} \in \mathbb{R}_+^{|\mathcal{A}|} \text{ s.t. } \quad \mathbf{x} = \sum_{\mathbf{w} \in \mathcal{W}} \mathbf{x}^{\mathbf{w}}, \quad \mathbf{N}\mathbf{x}^{\mathbf{w}} = \mathbf{d}^{\mathbf{w}} \quad \forall \mathbf{w} \in \mathcal{W} \right\}.$$

Let $c_a : \mathbb{R}_+^{|\mathcal{A}|} \to \mathbb{R}_+$ be the "cost" function for arc $a \in \mathcal{A}$. The interpretation of cost, here, is deliberately vague. The cost function might represent the actual time it takes to travel an arc, tolls users incur along that arc, disutility from environmental factors along that arc, or some combination of the above. Note that because of interdependencies in the network, the cost of traveling arc a may depend not only on $\mathbf{x}_a$, but on the flows on other arcs as well. Denote by $\mathbf{c}(\cdot)$ the vector-valued function whose a-th component is $c_a(\cdot)$.

A feasible flow $\mathbf{x}^*$ is a Wardrop equilibrium if for every origin–destination pair $\mathbf{w} \in W$, and any path connecting (w_s, w_t) with positive flow in $\mathbf{x}^*$, the cost of traveling along that path is less than or equal to the cost of traveling along any other path that connects (w_s, w_t). Here, the cost of traveling along a path is the sum of the costs of each of its constituent arcs. Intuitively, a Wardrop equilibrium captures the idea that if there exists a less congested route connecting w_s and w_t, users would find and use it instead of their current route.

It is well-known that a Wardrop equilibrium is a solution to VI$(\mathbf{c}, \mathcal{F})$.

2.2 Approximate equilibria

Let $\epsilon > 0$. We will say that $\hat{\mathbf{x}} \in \mathcal{F}$ is an ϵ-approximate solution to VI($\mathbf{f}, \mathcal{F}$) if

$$\mathbf{f}(\hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}}) \geq -\epsilon, \quad \forall \mathbf{x} \in \mathcal{F}. \tag{7}$$

This notion of an approximate solution is not new to the VI literature—it corresponds exactly to the condition that the primal gap function of the VI is bounded above by ϵ and is frequently used in the analysis of numerical procedures for solving the VI. We point out that ϵ-approximate solutions also frequently have a modeling interpretation. For example, consider the case of constrained convex optimization [cf. Eq. (2)]. Let $\mathbf{x}^*$ be an optimal solution. Since F is convex, we have $F(\hat{\mathbf{x}}) - F(\mathbf{x}^*) \leq -\nabla F(\hat{\mathbf{x}})^T(\mathbf{x}^* - \hat{\mathbf{x}}) \leq \epsilon$. In other words, ϵ-approximate solutions to VIs generalize the idea of ϵ-optimal solutions to convex optimization problems. Similarly, in a Nash equilibrium, an ϵ-approximate solution to the VI (5) describes the situation where each player i does not necessarily play her best response given what the other players are doing, but plays a strategy which is no worse than ϵ from her best response.

The idea of ϵ-approximate solutions is not the only notion of an approximate equilibrium. An alternative notion of approximation is that $\|\hat{\mathbf{x}} - \mathbf{x}^*\| \leq \delta$ where $\mathbf{x}^*$ is a solution to the VI($\mathbf{f}, \mathcal{F}$). We say such a $\hat{\mathbf{x}} \in \mathcal{F}$ is δ-near a solution to the VI($\mathbf{f}, \mathcal{F}$). As shown in Theorem 1, these two ideas are closely related. The theorem was proven in [33] to provide stopping criteria for certain types of iterative algorithms for solving VIs. We reinterpret it here in the context of approximate equilibria.

Before stating the theorem, we define strong monotonicity. We will say that $\mathbf{f}(\cdot)$ is *strongly monotone* if $\exists \gamma > 0$ such that

$$(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y}))^T(\mathbf{x} - \mathbf{y}) \geq \gamma \|\mathbf{x} - \mathbf{y}\|^2 \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{F}.$$

When the VI corresponds to constrained optimization [cf. Eqs. (2, 3)], strong monotonicity of f corresponds to strong convexity of F. Intuitively, strong monotonicity ensures that f does not have large, flat regions.

Theorem 1 ([33]) *Suppose* $\mathbf{f}$ *is strongly monotone with parameter* γ. *Then every* ϵ*-approximate solution to VI*($\mathbf{f}, \mathcal{F}$) *is* $\sqrt{\frac{\epsilon}{\gamma}}$*-near an exact solution.*

We require Theorem 1 in Sect. 7 to prove some of our generalization results.

2.3 Characterizing approximate solutions to VIs over conic representable sets

In this section we provide an alternative characterization of an ϵ-approximate solution [cf. Eq. (7)] in the case when $\mathcal{F}$ is represented by the intersection of conic inequalities.

Specifically, for the remainder of the paper, we will assume:

Assumption 1 $\mathcal{F}$ can be represented as the intersection of a small number of conic inequalities in standard form, $\mathcal{F} = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in C\}$.

Assumption 2 $\mathcal{F}$ satisfies a Slater-condition.

The assumption that $\mathcal{F}$ is given in standard form is not crucial. All of our results extend to the case that $\mathcal{F}$ is not given in standard form at the expense of some notation. It is, however, crucial, that $\mathcal{F}$ is conic representable. Observe that when C is the nonnegative orthant, we recover the special case where $\mathcal{F}$ is a polyhedron. With other choices of C, e.g., the second-order cone, we can model more complex sets, such as intersection of ellipsoids. To stress the dependence on $\mathbf{A}, \mathbf{b}, C$, we will write $\mathrm{VI}(\mathbf{f}, \mathbf{A}, \mathbf{b}, C)$.

The following result was first proven in [2] to describe a reformulation of $\mathrm{VI}(\mathbf{f}, \mathbf{A}, \mathbf{b}, C)$ as a single-level optimization problem. We reinterpret here as a characterization of approximate equilibria and sketch a short proof for completeness.

Theorem 2 ([2]) *Under assumptions* **A1**, **A2**, *the solution* $\hat{\mathbf{x}}$ *is an* ϵ*-approximate equilibrium to* $VI(\mathbf{f}, \mathbf{A}, \mathbf{b}, C)$ *if and only if* $\exists \mathbf{y}$ *s.t.*

$$\mathbf{A}^T\mathbf{y} \leq_C \mathbf{f}(\hat{\mathbf{x}}), \tag{8}$$

$$\mathbf{f}(\hat{\mathbf{x}})^T\hat{\mathbf{x}} - \mathbf{b}^T\mathbf{y} \leq \epsilon. \tag{9}$$

Proof First suppose that $\hat{\mathbf{x}}$ is an ϵ-approximate equilibrium. Then, from Eq. (7),

$$\mathbf{f}(\hat{\mathbf{x}})^T\hat{\mathbf{x}} - \epsilon \leq \mathbf{f}(\hat{\mathbf{x}})^T\mathbf{x}, \quad \forall \mathbf{x} \in \mathcal{F},$$

which is equivalent to $\mathbf{f}(\hat{\mathbf{x}})^T\hat{\mathbf{x}} - \epsilon \leq \min_{\mathbf{x}\in\mathcal{F}} \mathbf{f}(\hat{\mathbf{x}})^T\mathbf{x}$. The right hand side is a conic optimization problem in $\mathbf{x}$, and the above shows it is bounded below. Since $\mathcal{F}$ has non-empty interior, strong duality holds (see [12]), which implies that there exists a dual solution $\mathbf{y}$ that attains the optimum. In other words,

$$\min_{\mathbf{x}\in\mathcal{F}} \mathbf{f}(\hat{\mathbf{x}})^T\mathbf{x} = \max_{\mathbf{y}:\mathbf{A}^T\mathbf{y}\leq_C \mathbf{f}(\hat{\mathbf{x}})} \mathbf{b}^T\mathbf{y}.$$

Substituting this dual solution into the above inequality and rearranging terms yields the result. The reverse direction is proven analogously using weak conic duality. □

The above proof leverages the fact that the duality gap between an optimal primal and dual solution pair is zero. We can instead formulate a slightly different characterization by leveraging complementary slackness. In this case, Eq. (9) is replaced by the additional constraints

$$\sum_{i=1}^{n} x_i(f_i(\hat{\mathbf{x}}) - \mathbf{y}^T\mathbf{A}\mathbf{e}_i) \leq \epsilon. \tag{10}$$

Depending on the application, either the strong duality representation [cf. Eqs. (8, 9)] or the complementary slackness representation [cf. Eqs. (8, 10)] may be more natural. We will use the strong duality formulation in Sect. 8.3 and the the complementary slackness formulation in Sect. 8.1.

3 The inverse variational inequality problem

3.1 Problem formulation

We are now in a position to pose the inverse variational inequality problem. We are given observations $(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ for $j = 1, \ldots, N$. In this context, we modify Assumption **A2** to read.

Assumption The set $\mathcal{F}_j = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}_j\mathbf{x} = \mathbf{b}_j, \mathbf{x} \in C_j\}$ is non-empty and satisfies a Slater condition for each j.

This is not a particularly stringent condition; given data that does not satisfy it, we can always pre-process the data ensure it does satisfy this assumption.

We seek a function $\mathbf{f}$ such that $\mathbf{x}_j$ is an approximate solution to VI$(\mathbf{f}, \mathbf{A}_j, \mathbf{b}_j, C_j)$ for each j. Note, the function $\mathbf{f}$ is common to all observations. Specifically, we would like to solve:

$$\begin{aligned} \min_{\mathbf{f}, \boldsymbol{\epsilon}} \quad & \|\boldsymbol{\epsilon}\| \\ \text{s.t.} \quad & \mathbf{x}_j \text{ is an } \epsilon_j\text{-approximate solution to } \text{VI}(\mathbf{f}, \mathbf{A}_j, \mathbf{b}_j, C_j), \quad j = 1, \ldots, N, \\ & \mathbf{f} \in \mathcal{S}. \end{aligned} \tag{11}$$

where $\|\cdot\|$ represents some choice of norm, and $\mathcal{S}$ represents the set of admissible functions. In the parametric case, treated in the following section, we will assume that $\mathcal{S}$ is indexed by a vector of parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^M$. In the nonparametric case, $\mathcal{S}$ will be a general set of functions that satisfy certain smoothness properties. We defer this extension until Sect. 5.

3.2 Parametric estimation

In this section, we assume that the function $\mathbf{f}$ is known to belong to a parametric family indexed by a vector $\boldsymbol{\theta} \in \Theta$. We write $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$ to denote this dependence. We will assume throughout that Θ is compact and $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. A direct application of Theorem 2 yields the following reformulation:

Theorem 3 *Under assumptions 1, 2 and the additional constraint that* $\mathbf{f} = \mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$ *for some* $\boldsymbol{\theta} \in \Theta$*, problem Eq.* (11) *can be reformulated as*

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \Theta, \mathbf{y}, \boldsymbol{\epsilon}} \quad & \|\boldsymbol{\epsilon}\| \\ \textit{s.t.} \quad & \mathbf{A}_j^T \mathbf{y}_j \leq_C \mathbf{f}(\mathbf{x}_j; \boldsymbol{\theta}), \quad j = 1, \ldots N, \\ & \mathbf{f}(\mathbf{x}_j; \boldsymbol{\theta})^T \mathbf{x}_j - \mathbf{b}_j^T \mathbf{y}_j \leq \epsilon_j, \quad j = 1, \ldots, N, \end{aligned} \tag{12}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$.

Remark 1 (Multiple equilibria) We stress that since Theorem 2 is true for any ϵ-approximate solution to the VI, Theorem 3 is valid even when the function $\mathbf{f}$ might give rise to multiple distinct equilibria. This robustness to multiple equilibria is an important strength of our approach that distinguishes it from other specialized approaches that require uniqueness of the equilibrium.

Remark 2 (Equilibria on the boundary) In Theorem 2, we did not need to assume that the $\mathbf{x}_j$ or the solutions to $VI(\mathbf{f}, \mathbf{A}_j, \mathbf{b}_j)$ belonged to the interior of $\mathcal{F}_j$. Consequently, Theorem 3 is valid even if the observations $\mathbf{x}_j$ or induced solutions to $VI(\mathbf{f}, \mathbf{A}_j, \mathbf{b}_j)$ occur on the boundary. This is in contrast to many other techniques which require that the solutions occur on the relative interior of the feasible set.

Remark 3 (Computational complexity) Observe that $\mathbf{x}_j$ are data in Problem (12), not decision variables. Consequently, the complexity of this optimization depends on the cone C and the dependence of $\mathbf{f}$ on $\boldsymbol{\theta}$, but *not* on the dependence of $\mathbf{f}$ on $\mathbf{x}$. For a number of interesting parametric forms, we can show that Problem (12) is in fact tractable.

As an example, suppose $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i=1}^{M} \theta_i \boldsymbol{\phi}_i(\mathbf{x})$ where $\boldsymbol{\phi}_1(\mathbf{x}), \ldots, \boldsymbol{\phi}_M(\mathbf{x})$ is a set of (nonlinear) basis functions. Since $\mathbf{f}$ depends linearly on $\boldsymbol{\theta}$, Problem (12) is a conic optimization problem, even though the basis functions $\phi_i(\mathbf{x})$ may be arbitrary nonlinear functions. Indeed, if C is the nonnegative orthant, Problem (12) is a linear optimization problem. Similarly, if C is the second-order cone, Problem (12) is a second-order cone problem.

Finally, although structural estimation is not the focus of our paper, in Appendix 2 we briefly illustrate how to use Theorem 2 to formulate an alternate optimization problem that is similar to, but different from, Problem (12) and closer in spirit to structural estimation techniques. Moreover, we show that this formulation is equivalent to certain structural estimation techniques in the sense that they produce the same estimators. This section may prove useful to readers interested in comparing these methodology.

3.3 Application: demand estimation under Bertrand–Nash competition

In this section, we use Theorem 3 to estimate an unknown demand function for a product so that observed prices are approximately in Bertrand–Nash equilibrium. This is a somewhat stylized example inspired by various influential works in the econometrics literature, such as [9] and [10]. We include this stylized example for two reasons: 1) To illustrate a simple problem where equilibria may occur on the boundary of the feasible region. 2) To further clarify how the choice of parameterization of $\mathbf{f}(\cdot; \boldsymbol{\theta})$ affects the computational complexity of the estimation problem.

For simplicity, consider two firms competing by setting prices p_1, p_2, respectively. Demand for firm i's product, denoted $D_i(p_1, p_2, \xi)$, is a function of both prices, and other economic indicators, such as GDP, denoted by ξ. Each firm sets prices to maximize its own revenues $U_i(p_1, p_2, \xi) = p_i D_i(p_1, p_2, \xi)$ subject to the constraint $0 \leq p_i \leq \overline{p}$. The upper bound $\overline{p}$ might be interpreted as a government regulation as is frequent in some markets for public goods, like electricity. We assume a priori that each demand function belongs to some given parametric family indexed by $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta$: $D_1(p_1, p_2, \xi; \boldsymbol{\theta}_1)$, $D_2(p_1, p_2, \xi; \boldsymbol{\theta}_2)$. We seek to estimate $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ so that the data (p_1^j, p_2^j, ξ) for $j = 1, \ldots, N$ correspond approximately to Nash equilibria.

Both [9] and [10] assume that equilibrium prices do not occur on the boundary, i.e., that $p_i < \overline{p}$ since they leverage Eq. (6) in their analysis. These methods are, thus, not directly applicable.

By contrast, Theorem 3 directly applies yielding (after some arithmetic)

$$\begin{aligned}
&\min_{\substack{\mathbf{y},\boldsymbol{\epsilon} \\ (\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta}} \quad \|\boldsymbol{\epsilon}\| \\
&\text{s.t. } \mathbf{y}^j \geq \mathbf{0}, \quad j = 1,\ldots,N, \\
&y_i^j \geq p_i^j \frac{\partial}{\partial p_i} D_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i) + D_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i), \quad i = 1,2, \;\; j = 1,\ldots,N, \\
&\sum_{i=1}^{2} \overline{p}^j y_i^j - (p_i^j)^2 \frac{\partial}{\partial p_i} D_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i) \\
&\quad - p_i^j D_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i) \leq \epsilon_j, \quad j = 1,\ldots,N.
\end{aligned} \tag{13}$$

We stress that potentially more complex constraints on the feasible region can be incorporated just as easily.

Next, recall that the complexity of the optimization problem (13) depends on the parameterization of $D_i(p_1, p_2, \xi, \boldsymbol{\theta}_i)$. For example, when demand is linear,

$$D_i(p_1, p_2, \xi; \boldsymbol{\theta}_i) = \theta_{i0} + \theta_{i1} p_1 + \theta_{i2} p_2 + \theta_{i3} \xi \tag{14}$$

problem (13) reduces to the linear optimization problem:

$$\begin{aligned}
\min_{\mathbf{y},\boldsymbol{\epsilon},(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\in\Theta,\mathbf{d}} \quad & \|\boldsymbol{\epsilon}\| \\
\text{s.t.} \quad & \mathbf{y}^j \geq \mathbf{0}, \quad j = 1,\ldots,N, \\
& y_i^j \geq d_i^j + \theta_{ii} p_i^j, \qquad i = 1,2, \quad j = 1,\ldots,N, \\
& \overline{p} \sum_{i=1}^{2} y_i^j - p_i^j d_i^j - (p_i^j)^2 \theta_{ii} \leq \epsilon_j, \quad j = 1,\ldots,N, \\
& d_i^j = \theta_{i0} + \theta_{i1} p_1^j + \theta_{i2} p_2^j + \theta_{i3} \xi^j, \qquad i = 1,2, \quad j = 1,\ldots,N.
\end{aligned} \tag{15}$$

Alternatively, if we assume demand is given by the multinomial logit model [23], $D_i(p_1, p_2, \xi; \boldsymbol{\theta}) = \frac{e^{\theta_{i0}+\theta_{i1}p_i+\theta_{i3}\xi}}{e^{\theta_{10}+\theta_{11}p_1+\theta_{13}\xi} + e^{\theta_{20}+\theta_{21}p_2+\theta_{23}\xi} + \theta_{00}}$, the problem (13) becomes

$$\begin{aligned}
\min_{\mathbf{y},\boldsymbol{\epsilon},\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\mathbf{d_1},\mathbf{d_2}} \quad & \|\boldsymbol{\epsilon}\| \\
\text{s.t.} \quad & \mathbf{y}^j \geq \mathbf{0}, \quad j = 1,\ldots,N, \\
& y_i^j \geq p_i^j \theta_{i1} d_1^j d_2^j + d_i^j, \quad i = 1,2, \\
& \sum_{i=1}^{2} \overline{p}^j y_i^j + p_i^j d_i^j - (p_i^j)^2 \theta_{1i} d_i^j (1 - d_i^j) \leq \epsilon_j
\end{aligned}$$

$$d_i^j = \frac{e^{\theta_{0i} + \theta_{i1} p_i^j + \theta_{i3} \xi^j}}{e^{\theta_{10} + \theta_{11} p_1^j + \theta_{13} \xi^j} + e^{\theta_{20} + \theta_{21} p_2^j + \theta_{23} \xi} + \theta_{00}}, \; i = 1, 2, \; j = 1, \ldots N,$$

which is non-convex. Non-convex optimization problems can be challenging numerically and may scale poorly.

Finally, we point out that although it more common in the econometrics literature to specify the demand functions D_i directly as we have above, one could equivalently specify the marginal revenue functions

$$M_i(p_1, p_2, \xi; \boldsymbol{\theta}_i) = p_i \partial_i D_i(p_1, p_2, \xi; \boldsymbol{\theta}_i) + D_i(p_1, p_2, \xi; \boldsymbol{\theta}_i)$$

and then impute the demand function as necessary. We adopt this equivalent approach later in Sect. 8.1.

4 Kernel methods: background

Intuitively, our nonparametric approach in the next section seeks the "smoothest" function $\mathbf{f}$ which make the observed data approximate equilibria, where the precise notion of smoothness is determined by the choice of kernel. Kernel methods have been used extensively in machine learning, most recently for feature extraction in context of support-vector machines or principal component analysis. Our use of kernels, however, more closely resembles their application in spline interpolation and regularization networks ([24,39]).

Our goal in this section is to develop a sufficiently rich set of scalar valued functions over which we can tractably optimize using kernel methods. Consequently, we first develop some background. Our review is not comprehensive. A more thorough treatment of kernel methods can be found in either [36,38] or [21].

Let $\mathcal{F} \subseteq \mathbb{R}^n$ denote some domain. Let $k : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ be a symmetric function. We will say that k is a kernel if k is positive semidefinite over $\mathcal{F}$, i.e., if

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \text{ for any choice of } N \in \mathbb{N}, \;\; \mathbf{c} \in \mathbb{R}^N, \;\; \mathbf{x}_i \in \mathcal{F}.$$

Examples of kernels over $\mathbb{R}^n$ include:

Linear: $k(\mathbf{x}, \mathbf{y}) \equiv \mathbf{x}^T \mathbf{y}$,
Polynomial: $k(\mathbf{x}, \mathbf{y}) \equiv (c + \mathbf{x}^T \mathbf{y})^d$ for some choice of $c \geq 0$ and $d \in \mathbb{N}$,
Gaussian: $k(\mathbf{x}, \mathbf{y}) \equiv \exp(-c\|\mathbf{x} - \mathbf{y}\|^2)$ for some choice of $c > 0$.

Let $k_{\mathbf{x}}(\cdot) \equiv k(\mathbf{x}, \cdot)$ denote the function of one variable obtained by fixing the first argument of k to $\mathbf{x}$ for any $\mathbf{x} \in \mathcal{F}$. Define $\mathcal{H}_0$ to be the vector space of scalar valued functions which are representable as finite linear combinations of elements $k_{\mathbf{x}}$ for some $\mathbf{x} \in \mathcal{F}$, i.e.,

$$\mathcal{H}_0 = \left\{ \sum_{j=1}^{N} \alpha_j k_{\mathbf{x}_j} : \mathbf{x}_j \in \mathcal{F},\ N \in \mathbb{N},\ \alpha_j \in \mathbb{R},\ j = 1, \ldots, N, N \in \mathbb{N} \right\}. \tag{16}$$

Observe that $k_{\mathbf{x}} \in \mathcal{H}_0$ for all $\mathbf{x} \in \mathcal{F}$, so that in a sense these elements form a basis of the space $\mathcal{H}_0$. On the other hand, for a given $f \in \mathcal{H}_0$, its representation in terms of these elements $k_{\mathbf{x}_j}$ for $\mathbf{x}_j \in \mathcal{F}$ need not be unique. In this sense, the elements $k_{\mathbf{x}}$ are not like a basis.

For any $f, g \in \mathcal{H}_0$ such that

$$f = \sum_{j=1}^{N} \alpha_j k_{\mathbf{x}_j}, \quad g = \sum_{i=1}^{N} \beta_i k_{\mathbf{x}_i}, \quad \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^N \tag{17}$$

we define a scalar product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j \langle k_{\mathbf{x}_i}, k_{\mathbf{x}_j} \rangle_{\mathcal{H}_0} \equiv \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j). \tag{18}$$

Since the representation in (17) is not unique, for this to be a valid definition one must prove that the right-hand side of the last equality is independent of the choice of representation. It is possible to do so. See [36] for the details. Finally, given this scalar-product, we define the norm $\|f\|_{\mathcal{H}_0} \equiv \sqrt{\langle f, f \rangle_{\mathcal{H}_0}}$.

In what follows, we will actually be interested in the closure of $\mathcal{H}_0$, i.e.,

$$\mathcal{H} = \overline{\mathcal{H}_0}. \tag{19}$$

We extend the scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ and norm $\| \cdot \|_{\mathcal{H}_0}$ to $\mathcal{H}$ by continuity. (Again, see [36] for the details). Working with $\mathcal{H}$ instead of $\mathcal{H}_0$ simplifies many results.[1]

As an example, in the case of the linear and polynomial kernels, the space $\mathcal{H}$ is finite dimensional and corresponds to the space of linear functions and the space of polynomials of degree at most d, respectively. In the case of the Gaussian kernel, the space $\mathcal{H}$ is infinite dimensional and is a subspace of all continuous functions.

If $f \in \mathcal{H}_0$ admits a finite representation as in Eq. (17), note that from Eq. (18) we have for all $\mathbf{x} \in \mathcal{F}$

$$\langle k_{\mathbf{x}}, f \rangle_{\mathcal{H}} = \sum_{j=1}^{N} \alpha_j k(\mathbf{x}, \mathbf{x}_j) = f(\mathbf{x}). \tag{20}$$

In fact, it can be shown that this property applies to all $f \in \mathcal{H}$ ([30]). This is the most fundamental property of $\mathcal{H}$ as it allows us to relate the scalar product of the space to function evaluation. Equation (20) is termed the *reproducing property* and as a consequence, $\mathcal{H}$ is called a *Reproducing Kernel Hilbert Space (RKHS)*.

[1] For the avoidance of doubt, the closure in (19) is with respect to the norm $\| \cdot \|_{\mathcal{H}_0}$.

At this point, it may appear that RKHS are very restrictive spaces of functions. In fact, it can be shown that any Hilbert space of scalar-valued functions for which there exists a $c \in \mathbb{R}$ such that for each $f \in \mathcal{H}$, $|f(\mathbf{x})| \leq c\|f\|_{\mathcal{H}}$ for all $\mathbf{x} \in \mathcal{F}$ is an RKHS ([30]). Thus, RKHS are fairly general. Practically speaking, though, our three previous examples of kernels –linear, polynomial, and Gaussian –are by far the most common in the literature.

We conclude this section with a discussion about the norm $\|f\|_{\mathcal{H}}$. We claim that in each of our previous examples, the norm $\|f\|_{\mathcal{H}}$ makes precise a different notion of "smoothness" of the function f. For example, it is not hard to see that if $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, then under the linear kernel $\|f\|_{\mathcal{H}} = \|\mathbf{w}\|$. Thus, functions with small norm have small gradients and are "smooth" in the sense that they do not change value rapidly in a small neighborhood.

Similarly, it can be shown (see [24]) that under the Gaussian kernel,

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{(2\pi)^n} \int |\tilde{f}(\omega)|^2 e^{\frac{\|\omega\|^2}{2c}} \, d\omega, \tag{21}$$

where $\tilde{f}$ is the Fourier transformation of f. Thus, functions with small norms do not have many high-frequency Fourier coefficients and are "smooth" in the sense that they do not oscillate very quickly.

The case of the polynomial kernel is somewhat more involved as there does not exist a *simple* explicit expression for the norm (see [24]). However, it is easily confirmed numerically using Eq. (18) that functions with small norms do not have large coefficients and do not have have high degree. Consequently, they are "smooth" in the sense that their derivatives do not change value rapidly in a small neighborhood.

Although the above reasoning is somewhat heuristic, it is possible to make the intuition that the norm on an RKHS describes a notion of smoothness completely formal. The theoretical details go beyond the scope of this paper (see [24]). For our purposes, an intuitive appreciation that the $\mathcal{H}$-norm penalizes non-smooth functions and that the particular notion of smoothness is defined by the kernel will be sufficient for the remainder.

5 The inverse variational inequality problem: a nonparametric approach

5.1 Kernel based formulation

In this section, we develop a nonparametric approach to the inverse variational inequality problem. The principal difficulty in formulating a nonparametric equivalent to (11) is that the problem is ill-posed. Specifically, if the set S is sufficiently rich, we expect there to be many, potentially infinitely many, different functions $\mathbf{f}$ which all reconcile the data, and make each observation an exact equilibrium. Intuitively, this multiplicity of solutions is similar to the case of interpolation where, given a small set of points, many different functions will interpolate between them exactly. Which function, then, is the "right" one?

We propose to select the function $\mathbf{f}$ of minimal $\mathcal{H}$-norm among those that approximately reconcile the data. This choice has several advantages. First, as mentioned earlier, functions with small norm are "smooth", where the precise definition of smoothness will be determined by the choice of kernel. We feel that in many applications, assuming that the function defining a VI is smooth is very natural. Second, as we shall prove, identifying the function $\mathbf{f}$ with minimal norm is computationally tractable, even when the RKHS $\mathcal{H}$ is infinite dimensional. Finally, as we will show in Sect. 7, functions with bounded $\mathcal{H}$-norm will have good generalization properties.

Using Theorem 2, we reformulate Problem (11) as

$$\begin{aligned}
\min_{\mathbf{f},\mathbf{y},\boldsymbol{\epsilon}} \quad & \sum_{i=1}^{n} \| f_i \|_{\mathcal{H}}^2 \\
\text{s.t.} \quad & \mathbf{A}_j^T \mathbf{y}_j \leq \mathbf{f}(\mathbf{x}_j), \qquad j = 1, \ldots, N, && \text{(22a)} \\
& \mathbf{x}_j^T \mathbf{f}(\mathbf{x}_j) - \mathbf{b}_j^T \mathbf{y}_j \leq \epsilon_j, \quad j = 1, \ldots, N, && \text{(22b)} \\
& \|\boldsymbol{\epsilon}\| \leq \kappa, \quad \boldsymbol{\epsilon} \geq \mathbf{0}, \quad f_i \in \mathcal{H}, \quad i = 1, \ldots, n, \\
& \frac{1}{N} \sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{f}(\mathbf{x}_j) = 1. && \text{(22c)}
\end{aligned}$$

Here f_i is the i-th component of the vector function $\mathbf{f}$ and $\mathcal{H}$ is an RKHS. Since we may always scale the function $\mathbf{f}$ in VI($\mathbf{f}$, $\mathcal{F}$) by a positive constant without affecting the solution, we require the last constraint as a normalization condition. Finally, the exogenous parameter κ allows us to balance the norm of $\mathbf{f}$ against how closely $\mathbf{f}$ reconciles the data; decreasing κ will make the observed data closer to equilibria at the price of $\mathbf{f}$ having greater norm.

Problem (22) is an optimization over functions, and it is not obvious how to solve it. We show in the next theorem, however, that this can be done in a tractable way. This theorem is an extension of a representation theorem from the kernel literature (see [39]) to the constrained multivariate case. See the Appendix for a proof.

Theorem 4 *Suppose Problem* (22) *is feasible. Then, there exists an optimal solution* $\mathbf{f}^* = (f_1^*, \ldots, f_n^*)$ *with the following form:*

$$f_i^* = \sum_{j=1}^{N} \alpha_{i,j} k_{\mathbf{x}_j}, \tag{23}$$

for some $\alpha_{i,j} \in \mathbb{R}$, *where k denotes the kernel of* $\mathcal{H}$.

By definition of $\mathcal{H}$, when Problem (22) is feasible, its solution is a potentially infinite expansion in terms of the kernel function evaluated at various points of $\mathcal{F}$. The importance of Theorem 4 is that it allows us to conclude, first, that this expansion is in fact finite, and second, that the relevant points of evaluation are exactly the data points $\mathbf{x}_j$. This fact further allows us to replace the optimization problem (22), which is over an infinite dimensional space, with an optimization problem over a finite dimensional space.

Theorem 5 *Problem (22) is feasible if and only if the following optimization problem is feasible:*

$$\begin{aligned} \min_{\boldsymbol{\alpha},\mathbf{y},\boldsymbol{\epsilon}} \quad & \sum_{i=1}^{n} \mathbf{e}_i^T \boldsymbol{\alpha}\mathbf{K}\boldsymbol{\alpha}^T \mathbf{e}_i \\ \text{s.t.} \quad & \mathbf{A}_j \mathbf{y}_j \le \boldsymbol{\alpha}\mathbf{K}\mathbf{e}_j \quad j = 1, \ldots, N, \\ & \mathbf{x}_j^T \boldsymbol{\alpha}\mathbf{K}\mathbf{e}_j - \mathbf{b}_j^T \mathbf{y}_j \le \epsilon_j \quad j = 1, \ldots, N, \\ & \|\boldsymbol{\epsilon}\| \le \kappa, \quad \boldsymbol{\epsilon} \ge \mathbf{0}, \\ & \frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_j^T \boldsymbol{\alpha}\mathbf{K}\mathbf{e}_j = 1. \end{aligned} \tag{24}$$

Here $\boldsymbol{\alpha} = (\alpha_{ij})_{i=1,j=1}^{i=n,j=N} \in \mathbb{R}^{n\times N}$, and $\mathbf{K} = (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^{i,j=N}$. Moreover, given an optimal solution $\boldsymbol{\alpha}$ to the above optimization problem, an optimal solution to Problem (22) is given by Eq. (23).

See the Appendix for a proof. Given the optimal parameters $\boldsymbol{\alpha}$, $\mathbf{f}$ can be evaluated at new points $\mathbf{t}$ using (23). Note that $\mathbf{K}$ is positive semidefinite (as a matrix) since k is positive definite (as a function). Thus, (24) is a convex, quadratic optimization problem. Such optimization problems are very tractable numerically and theoretically, even for large-scale instances. (See [12]). Moreover, this quadratic optimization problem exhibits block structure—only the variables $\boldsymbol{\alpha}_j$ couple the subproblems defined by the $\mathbf{y}_j$—which can be further exploited in large-scale instances. Finally, the size of this optimization scales with N, the number of observations, not with the dimension of the original space $\mathcal{H}$, which may be infinite.

Observe that Problem (22) is bounded, but may be infeasible. We claim it will be feasible whenever κ is sufficiently large. Indeed, let $\hat{\mathbf{f}}_i \in \mathcal{H}$ be any functions from the RKHS. By scaling, we can always ensure (22c) is satisfied. The following convex optimization $\min_{\mathbf{x}:\mathbf{A}_j\mathbf{x}=\mathbf{b}_j,\mathbf{x}\ge\mathbf{0}} \hat{\mathbf{f}}(\mathbf{x}_j)^T\mathbf{x}$ is bounded and satisfies a Slater condition by Assumption **A2**. Let $\hat{\mathbf{y}}_j$ be the dual variables to this optimization, so that $\hat{\mathbf{y}}_j$ satisfy (22a) and define $\hat{\epsilon}_j$ according to (22b). Then as long as $\kappa \ge \|\hat{\boldsymbol{\epsilon}}\|$, Problem (22), and consequently Problem (24), will be feasible and obtain an optimal solution.

Computationally, treating the possible infeasibility of (24) can be cumbersome, so in what follows, we find it more convenient to dualize this constraint so that the objective becomes,

$$\min_{\boldsymbol{\alpha},\mathbf{y}} \boldsymbol{\alpha}^T K \boldsymbol{\alpha} + \lambda\|\boldsymbol{\epsilon}\|, \tag{25}$$

and then solve this problem for various choices of $\lambda > 0$. Note this version of the problem is always feasible, and, indeed, we will employ this formulation later in Sect. 8.

We conclude this section by contrasting our parametric and nonparametric formulations. Unlike the parametric approach, the nonparametric approach is always a convex optimization problem. This highlights a key tradeoff in the two approaches. The parametric approach offers us fine-grained control over the specific form of the function

$\mathbf{f}$ at the potential expense of the tractability of the optimization. The nonparametric approach offers less control but is more tractable.

We next illustrate our nonparametric approach below with an example.

5.2 Application: estimating the cost function in Wardrop equilibrium

Recall the example of Wardrop equilibrium from Sect. 2.1. In practice, while the network $(\mathcal{V}, \mathcal{A})$ is readily observable, the demands $\mathbf{d}^w$ and cost function $c_a(\cdot)$ must be estimated. Although several techniques already exist for estimating the demands $\mathbf{d}^w$ ([1,40]), there are fewer approaches for estimating $c_a(\cdot)$. Those techniques that do exist often use stylized networks, e.g., one origin–destination pair, to build insights. See [31] for a maximum likelihood approach, and [34] for kinematic wave analyses.

By contrast, we focus on estimating $c_a(\cdot)$ from observed flows or traffic counts on real, large scale networks. Specifically, we assume we are given networks $(\mathcal{V}_j, \mathcal{A}_j)$, $j = 1, \ldots, N$, and have access to estimated demands on these networks $\mathbf{d}^{\mathbf{w}_j}$ for all $\mathbf{w}_j \in W_j$. In practice, this may be the same network observed at different times of day, or different times of year, causing each observation to have different demands.

In the transportation literature, one typically assumes that $c_a(\cdot)$ only depends on arc a, and in fact, can be written in the form $c_a(x_a) = c_{0a} g\left(\frac{x_a}{m_a}\right)$, for some nondecreasing function g. The constant c_{0a} is sometimes called the free-flow travel time of the arc, and m_a is the effective capacity of the arc. These constants are computed from particular characteristics of the arc, such as its length, the number of lanes or the posted speed limit. (Note the capacity m_a is not a hard constraint; it not unusual to see arcs where $x_a^* > m_a$ in equilibrium.) We will also assume this form for the cost function, and seek to estimate the function $g(\cdot)$.

Using (24) and (25) we obtain the quadratic optimization problem

$$\min_{\boldsymbol{\alpha},\mathbf{y},\boldsymbol{\epsilon}} \quad \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} + \lambda \|\boldsymbol{\epsilon}\| \tag{26}$$

$$\begin{aligned}
\text{s.t.} \quad & \mathbf{e}_a^T \mathbf{N}_j^T \mathbf{y}^{\mathbf{w}} \le c_{0a} \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_a, \quad \forall \mathbf{w} \in W_j,\ a \in \mathcal{A}_j,\ j = 1, \ldots, N, \\
& \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_a \le \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_{a'}, \qquad \forall a, a' \in \mathcal{A}_0 \ \text{ s.t. } \frac{x_a}{m_a} \le \frac{x_{a'}}{m_{a'}}, \qquad (27) \\
& \sum_{a \in \mathcal{A}_j} c_{0a} x_a \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_a - \sum_{\mathbf{w} \in W_j} (\mathbf{d}^{\mathbf{w}})^T \mathbf{y}^{\mathbf{w}} \le \epsilon_j, \quad , \forall \mathbf{w} \in W_j,\ j = 1, \ldots, N, \\
& \boldsymbol{\alpha}^T \mathbf{K} \mathbf{e}_{a_0} = 1.
\end{aligned}$$

In the above formulation $\mathcal{A}_0$ is a subset of $\bigcup_{j=1}^N \mathcal{A}_j$ and $\mathbf{K} \in \mathbb{R}^{\sum_{j=1}^N |\mathcal{A}_j| \times \sum_{j=1}^N |\mathcal{A}_j|}$. Constraint (27) enforces that the function $g(\cdot)$ be non-decreasing on these arcs. Finally, a_0 is some (arbitrary) arc chosen to normalize the function.

Notice, the above optimization can be quite large. If the various networks are of similar size, the problem has $O(N(|\mathcal{A}_1| + |W_1||\mathcal{V}_1|)$ variables and $O(N|W_1||\mathcal{A}_1| + |\mathcal{A}_0|)$ constraints. As mentioned previously, however, this optimization exhibits significant structure. First, for many choices of kernel, the matrix K is typically (approximately) low-rank. Thus, it is usually possible to reformulate the optimization in a much lower dimensional space. At the same time, for a fixed value of α, the optimization decouples by $\mathbf{w} \in W_j$ and j. Each of these subproblems, in turn, is a shortest path problem which can be solved very efficiently, even for large-scale networks. Thus, combining an appropriate transformation of variables with block decomposition, we can solve fairly large instances of this problem. We take this approach in Sect. 8.3.

6 Extensions

Before proceeding, we note that Theorem 4 actually holds in a more general setting. Specifically, a minimization over an RKHS will admit a solution of the form (23) whenever

(a) the optimization only depends on the norms of the components $\|f_i\|_{\mathcal{H}}$ and the function evaluated at a finite set of points $\mathbf{f}(\mathbf{x}_j)$, and
(b) the objective is nondecreasing in the norms $\|f_i\|_{\mathcal{H}}$.

The proof is identical to the one presented above, and we omit it for conciseness. An important consequence is that we can leverage the finite representation of Theorem 4 in a number of other estimation problems and to facilitate inference. In this section, we describe some of these extensions.

6.1 Incorporating priors and semi-parametric estimation

Suppose we believe a priori that the function $\mathbf{f}$ describing the VI should be close to a particular function $\mathbf{f}_0$ (a prior). In other words, $\mathbf{f} = \mathbf{f}_0 + \mathbf{g}$ for some function $\mathbf{g}$ which we believe is small. We might then solve

$$\begin{aligned} \min_{\mathbf{g},\mathbf{y},\boldsymbol{\epsilon}} \quad & \sum_{i=1}^{n} \|g_i\|_{\mathcal{H}}^2 \\ \text{s.t.} \quad & \mathbf{A}_j^T \mathbf{y}_j \le \mathbf{f}_0(\mathbf{x}_j) + \mathbf{g}(\mathbf{x}_j) \qquad j = 1, \ldots, N, \\ & \mathbf{x}_j^T(\mathbf{f}_0(\mathbf{x}_j) + \mathbf{g}(\mathbf{x}_j)) - \mathbf{b}_j^T \mathbf{y}_j \le \epsilon_j \quad j = 1, \ldots, N, \\ & \|\boldsymbol{\epsilon}\| \le \kappa, \quad \boldsymbol{\epsilon} \ge \mathbf{0}, \quad g_i \in \mathcal{H}, \quad i = 1, \ldots, n. \end{aligned}$$

From our previous remarks, it follows that this optimization is equivalent to

$$\begin{aligned} \min_{\boldsymbol{\alpha},\mathbf{y},\boldsymbol{\epsilon}} \quad & \sum_{i=1}^{n} \mathbf{e}_i^T \boldsymbol{\alpha} \mathbf{K} \boldsymbol{\alpha}^T \mathbf{e}_i \\ \text{s.t.} \quad & \mathbf{A}_j \mathbf{y}_j \le \mathbf{f}_0(\mathbf{x}_j) + \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j \quad j = 1, \ldots, N \\ & \mathbf{x}_j^T(\mathbf{f}_0(\mathbf{x}_j) + \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j) - \mathbf{b}_j^T \mathbf{y}_j \le \epsilon_j \quad j = 1, \ldots, N, \\ & \|\boldsymbol{\epsilon}\| \le \kappa, \quad \boldsymbol{\epsilon} \ge \mathbf{0}, \end{aligned}$$

which is still a convex quadratic optimization problem.

In a similar way we can handle semi-parametric variants where $\mathbf{f}$ decomposes into the sum of two functions, one of which is known to belong to a parametric family and the other of which is defined nonparmetrically, i.e., $\mathbf{f}(\cdot) = \mathbf{f}_0(\cdot; \boldsymbol{\theta}) + \mathbf{g}$ for some $\boldsymbol{\theta}$ and $\mathbf{g} \in \mathcal{H}^n$.

Remark 4 (A Challenge with Partial Derivatives) There are natural modeling circumstances where Theorem 4 is not applicable. For example, recall in our demand estimation example from Sect. 3.3 that the inverse variational inequality problem depends not only on the demand functions $D_1(\cdot)$, $D_2(\cdot)$ evaluated at a finite set of points (p_1^j, p_2^j), but also on their partial derivatives at those points. Intuitively, the partial derivative $\partial_i D_i(p_1^j, p_2^j)$ requires information about the function in a small neighborhood of (p_1^j, p_2^j), not just at the point, itself. Consequently, Theorem 4 is not applicable. Extending the above techniques to this case remains an open area of research.

6.2 Ambiguity sets

In many applications, there may be multiple distinct models which all reconcile the data equally well. Breiman termed this phenomenon the "Rashomon" effect. It can occur even with parametric models that are well-identified, since there may exist models outside the parametric family which will also reconcile the data. Consequently, we would often like to identify the range of functions which may explain our data, and how much they differ.

We can determine this range by computing the upper and lower envelopes of the set of all functions within an RKHS that make the observed data approximate equilibria. We call this set the ambiguity set for the estimator. To construct these upper and lower bounds on the ambiguity set, consider fixing the value of κ in (22) and replacing the objective by $f_i(\hat{\mathbf{x}})$ for some $\hat{\mathbf{x}} \in \mathcal{F}$. This optimization problem satisfies the two conditions listed at the beginning of this section. Consequently, Theorem 4 applies, and we can use the finite representation to rewrite the optimization problem as a linear optimization problem in $\boldsymbol{\alpha}, \mathbf{y}$. Using software for linear optimization, it is possible to generate lower and upper bounds on the function $\mathbf{f}(\hat{\mathbf{x}})$ for various choices of $\hat{\mathbf{x}}$ quickly and efficiently.

To what value should we set the constant κ? One possibility is to let κ be the optimal objective value of (12), or a small multiple of it. This choice of κ yields the set of functions which "best" reconcile the given data. We discuss an alternative approach in Sect. 7 that yields a set of functions which are statistically similar to the current estimator.

Regardless of how we choose, κ, though, ambiguity sets can be combined with our previous parametric formulations to assess the appropriateness of the particular choice of parametric family. Indeed, the ambiguity set formed from the nonparametric kernel contains a set of alternatives to our parametric form which are, in some sense, equally plausible from the data. If these alternatives have significantly different behavior from our parametric choice, we should exercise caution when interpreting the fitted function.

Can we ever resolve the Rashomon effect? In some cases, we can use application-specific knowledge to identify a unique choice. In other cases, we need appeal to some extra, a priori criterion. A typical approach in machine learning is to focus on a choice with good generalizability properties. In the next section, we show that our proposed estimators enjoy such properties.

7 Generalization guarantees

In this section, we seek to prove generalization guarantees on the estimators from Problem (12) and (22). Proving various types of generalization guarantees for various algorithms is a central problem in machine learning. These guarantees ensure that the performance of our estimator on new, future data will be similar to its observed performance on existing data.

We impose a mild assumption on the generating process which is common throughout the machine learning literature:

Assumption 3 The data $(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ are i.i.d. realizations of random variables $(\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$ drawn from some probability measure $\mathbb{P}$.

Notice, we make no assumptions on potential dependence between $(\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$, nor do we need to know the precise form of $\mathbb{P}$. We also assume

Assumption 4 The random set $\tilde{\mathcal{F}} = \{\mathbf{x} : \tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}}, \mathbf{x} \in \tilde{C}\}$ satisfies a Slater Condition almost surely.

Assumption 5 $\tilde{\mathbf{x}} \in \tilde{\mathcal{F}}$ almost surely.

Assumptions 4 and 5 are not particularly stringent. If these condition may fail, we can consider pre-processing the data so that they succeed, and then consider a new measure $\mathbb{Q}$ induced by this processing of $\mathbb{P}$.

We now prove a bound for a special case of Problem (12). Let $z_N, \boldsymbol{\theta}_N$ denote the optimal value and optimal solution of (12). If for some N, there exist multiple optimal solutions, choose $\boldsymbol{\theta}_N$ by some tie-breaking rule, e.g., the optimal solution with minimal ℓ_2-norm. For any $0 < \alpha < 1$, define

$$\beta(\alpha) \equiv \sum_{i=0}^{\dim(\boldsymbol{\theta})} \binom{N}{i} \alpha^i (1-\alpha)^{N-i}.$$

Theorem 6 *Consider Problem* (12) *where the norm* $\|\cdot\| = \|\cdot\|_\infty$. *Suppose that this problem is convex in* $\boldsymbol{\theta}$ *and that Assumptions* **A1**, **A3**–**A5** *hold. Then, for any* $0 < \alpha < 1$, *with probability at least* $1 - \beta(\alpha)$ *with respect to the sampling,*

$$\mathbb{P}\left(\tilde{\mathbf{x}} \textit{ is a } z_N\textit{-approximate equilibrium for } VI(\mathbf{f}(\cdot, \boldsymbol{\theta}_N), \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})\right) \geq 1 - \alpha.$$

The proof relies on relating Problem (12) to an uncertain convex program [18], and leveraging results on the randomized counterparts of such programs. See the Appendix for the details.

Remark 5 There are two probability measures in the theorem. The first (explicit) is the probability measure of the new data point $(\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$. The second (implicit) is the probability measure of the random quantities $z_N, \boldsymbol{\theta}_N$. One way to interpret the theorem is as follows: One can ask, "For a fixed pair $z_N, \boldsymbol{\theta}_N$, is the probability that $\mathbf{x}_{N+1}$ is a z_N-approximate equilibrium for $\mathrm{VI}(\mathbf{f}(\cdot, \boldsymbol{\theta}_N), \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1})$ with respect to the first measure at least $1-\alpha$?" The theorem asserts the answer is, "Yes" with probability at least $1-\beta(\alpha)$ with respect to the second measure. More loosely, the theorem asserts that for "typical" values of $z_N, \boldsymbol{\theta}_N$, the answer is "yes." This type of generalization result, i.e., a result which is conditional on the data-sampling measure, is typical in machine learning.

Remark 6 Notice that $\beta(\alpha)$ corresponds to the tail probability of a binomial distribution, and, hence, converges exponentially fast in N.

Remark 7 (ℓ_1 Regularization) The value $\beta(\alpha)$ depends strongly on the dimension of θ. In [17], the authors show that including an ℓ_1 regularization of the form $\|\theta\|_1$ to reduce the effective dimension of θ can significantly improve the above bound in the context of uncertain convex programs.[2] Motivated by this idea, we propose modifying our original procedure by including a regularization $\lambda\|\boldsymbol{\theta}\|_1$ in the objective of Problem (12). Since the problem is convex this formulation is equivalent to including a constraint of the form $\|\boldsymbol{\theta}\|_1 \leq \kappa$ for some value of κ that implicitly depends on λ, and, consequently, Theorem 6 still applies but with z_N redefined to exclude the contribution of the regularization to the objective value.

Unfortunately, the proof of Theorem 6 doesn't generalize easily to other problems, such as other norms or Problem (22). A more general approach to proving generalization bounds is based upon Rademacher complexity. Rademacher complexity is a popular measure of the complexity of a class of functions, related to the perhaps better known VC-bounds. Loosely speaking, for function classes with small Rademacher complexity, empirical averages of functions in the class converge to their true expectation uniformly over the class, and there exist bounds on the rate of convergence which are tight up to constant factors. We refer the reader to [7] for a formal treatment.

We will use bounds based upon the Rademacher complexity of an appropriate class of functions to prove generalization bounds for both our parametric and nonparametric approaches. In the case of our nonparametric approach, however, it will prove easier to analyze the following optimization problem instead of Problem (22):

$$\begin{aligned} \min_{\mathbf{f},\mathbf{y},\boldsymbol{\epsilon}} \quad & \frac{\|\boldsymbol{\epsilon}\|_p^p}{N} \\ \text{s.t.} \quad & \mathbf{A}_j^T \mathbf{y}_j \leq \mathbf{f}(\mathbf{x}_j), \qquad j = 1, \ldots, N, \\ & \mathbf{x}_j^T \mathbf{f}(\mathbf{x}_j) - \mathbf{b}_j^T \mathbf{y}_j \leq \epsilon_j, \quad j = 1, \ldots, N, \end{aligned} \tag{28}$$

[2] In fact, the authors show more: they give an algorithm leveraging ℓ_1 regularization to reduce the dimensionality of θ and then an improved bound based on the reduced dimension. The analysis of this improved bound can be adapted to our current context at the expense of more notation. We omit the details for space.

$$\|f_i\|_{\mathcal{H}}^2 \leq \kappa_i, \quad f_i \in \mathcal{H}, \quad i = 1, \ldots, n,$$
$$\frac{1}{N}\sum_{j=1}^{N} \mathbf{x}_j^T \mathbf{f}(\mathbf{x}_j) = 1,$$

for some fixed p, $1 \leq p < \infty$. We have made two alterations from Problem (22) with the dualized objective (25). First, using Lagrangian duality, we have moved the term $\lambda \sum_{i=1}^n \|f_i\|_{\mathcal{H}}$ from the objective to the constraints. Indeed, for any value of λ, there exists values κ_i so that these two problems are equivalent. Second, we have specialized the choice of norm to a p-norm, and then made an increasing transformation of the objective. If we can show that solutions to Problem (28) enjoy strong generalization guarantees, Problem (22) should satisfy similar guarantees. Now, introduce

Assumption 6 The set $\tilde{\mathcal{F}}$ is contained within a ball of radius R almost surely.

Next, we introduce some additional notation. Consider Problem (12). Define

$$2 \sup_{\substack{\mathbf{x}:\|\mathbf{x}\|_2 \leq R \\ \boldsymbol{\theta} \in \Theta}} \|\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})\|_2 \equiv \overline{B}. \tag{29}$$

Observe $\overline{B} < \infty$. Let $\mathbf{f}_N = \mathbf{f}(\cdot; \boldsymbol{\theta}_N)$ denote the function corresponding to the optimal solution of Problem (12). With a slight abuse of language, we call $\mathbf{f}_N$ a solution to Problem (12).

We define analogous quantities for Problem (28). Given a kernel $k(\cdot, \cdot)$, let $\overline{K}^2 \equiv \sup_{\mathbf{x}:\|\mathbf{x}\|_2 \leq R} k(\mathbf{x}, \mathbf{x})$. Notice, if k is continuous, $\overline{K}$ is finite by **A6**. For example,

$$\overline{K}^2 = \begin{cases} R^2 & \text{for the linear kernel} \\ (c + R^2)^d & \text{for the polynomial kernel} \\ 1 & \text{for the Gaussian kernel} \end{cases} \tag{30}$$

With a slight abuse of notation, let $z_N, \mathbf{f}_N$ denote the optimal value and an optimal solution to Problem (28), and let $\overline{B} \equiv 2R\overline{K}\sqrt{\sum_{i=1}^n \kappa_i^2}$. This mild abuse of notation allows us to express our results in a unified manner. It will be clear from context whether we are treating Problem (12) or Problem (28), and consequently be clear which definition of $\mathbf{f}_N, \overline{B}$ we mean.

Finally, define $\epsilon(\mathbf{f}_N, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$ to be the smallest $\epsilon \geq 0$ such that $\tilde{\mathbf{x}}$ is an ϵ-approximate solution to $\mathrm{VI}(\mathbf{f}_N, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$.

Theorem 7 *Let $z_N, \mathbf{f}_N$ be the optimal objective and an optimal solution to Problem (12) or (28). Assume* **A1**, **A3**–**A6**. *For any $0 < \beta < 1$, with probability at least $1 - \beta$ with respect to the sampling,*

(i)

$$\mathbb{E}[(\epsilon(\mathbf{f}_N, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C}))^p] \leq z_N + \frac{1}{\sqrt{N}}\left(4p\overline{B}^p + 2\overline{B}^{p/2}\sqrt{2\log(2/\beta)}\right). \tag{31}$$

(ii) For any $\alpha > 0$,

$$\mathbb{P}(\tilde{\mathbf{x}} \text{ is a } z_N{+}\alpha\text{-approximate equilibrium for } \mathrm{VI}(\mathbf{f}_N, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})) \geq 1 - \frac{1}{\alpha^p \sqrt{N}} \left(4p\overline{B}^p + 2\overline{B}^{p/2} \sqrt{2 \log(2/\beta)} \right).$$

Remark 8 To build some intuition, consider the case $p = 1$. The quantity z_N is the average error on the data set for $\mathbf{f}_N$. The theorem shows with high-probability, $\mathbf{f}_N$ will make a new data point an ϵ-approximate equilibrium, where ϵ is only $O(1/\sqrt{N})$ larger than z_N. In other words, the fitted function will perform not much worse than the average error on the old data. Note, this does not guarantee that z_N is small. Indeed, z_N will only be small if in fact a VI is a good model for the system.

Remark 9 (Specifying Ambiguity Sets) We can use Theorem 7 to motivate an alternate proposal for specifying κ in ambiguity sets as in Sect. 6. Specifically, let R_N denote the second term on the righthand side of (31). Given another feasible function $\mathbf{f}'$ in Problem (28) whose objective value is strictly greater than $z_N + R_N$, we can claim that with probability at least $1 - \beta$, $\mathbf{f}_N$ has a smaller expected approximation error than $\mathbf{f}'$. However, if the objective value of $\mathbf{f}'$ is smaller than $z_N + R_N$, we cannot reject it at level $1 - \beta$; it is statistically as plausible as $\mathbf{f}_N$. Setting $\kappa = R_N$ in our ambiguity set recovers all such "statistically plausible" functions.

Theorems 6 and 7 provide a guarantee on the generalization error of our method. We may also be interested in its predictive power. Namely, given a new point $(\mathbf{x}_{N+1}, \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1})$, let $\hat{\mathbf{x}}$ be a solution to $\mathrm{VI}(\mathbf{f}_N, \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1}))$ The value $\hat{\mathbf{x}}$ is a prediction of the state of a system described by $(\mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1})$ using our fitted function and $\mathbf{x}_{N+1}$ represents true state of that system. We have the following theorem:

Theorem 8 *Assume $\mathbf{f}_N$ is strongly monotone with parameter γ.*

(i) Suppose the conditions of Theorem 6 hold. Then, for any $0 < \alpha < 1$, with probability at least $1 - \beta(\alpha)$ with respect to the sampling,

$$\|\mathbf{x}_{N+1} - \hat{\mathbf{x}}\| \leq \sqrt{\frac{z_N}{\gamma}}.$$

(ii) Suppose the conditions of Theorem 7 hold. Then, for any $0 < \beta < 1$, with probability at least $1 - \beta$ with respect to the sampling, for any $\alpha > 0$

$$\mathbb{P}\left(\|\mathbf{x}_{N+1} - \hat{\mathbf{x}}\| > \sqrt{\frac{z_N + \alpha}{\gamma}} \right) \leq \frac{1}{\alpha^p \sqrt{N}} \left(4p\overline{B}^p + 2\overline{B}^{p/2} \sqrt{2 \log(2/\beta)} \right).$$

In words, Theorem 8 asserts that solutions to our VI using our fitted function serve as good predictions to future data realizations. This is an important strength of our approach as it allows us to predict future behavior of the system. Again, this is contingent on the fact that z_N is small, i.e., that the VI well-explains the current data.

We conclude this section by noting that experimental evidence from machine learning suggests that bounds such as those above based on Rademacher complexity can be loose in small-samples. The recommended remedy is that, when computationally feasible, to use a more numerically intensive method like cross-validation or bootstrapping to estimate approximation and prediction errors. This approach applies equally well to choosing parameters like the threshold in an ambiguity set κ as described in Remark 9. We employ both approaches in Sect. 8.

8 Computational experiments

In this section, we provide some computational experiments illustrating our approach. For concreteness, we focus on our two previous examples: estimating the demand function in Bertrand–Nash equilibrium from Sect. 3.3 and estimating cost functions in traffic equilibrium from Sect. 5.2.

Before providing the details of the experiments, we summarize our major insights.

1. In settings where there are potentially many distinct functions that explain the data equally well, our nonparametric ambiguity sets are able to identify this set of functions. By contrast, parametric methods may misleadingly suggest there is only one possible function.
2. Even in the presence of endogenous, correlated noise, our parametric and nonparametric techniques are able to learn functions with good generalizability, even if the specified class does not contain the true function generating the data.
3. Sometimes, the functions obtained by our method are not strongly monotone. Nonetheless, they frequently still have reasonable predictive power.

8.1 Bertrand–Nash equilibrium (full-information)

We first consider an idealized, full-information setting to illustrate the importance of our ambiguity set technique. Specifically, we assume the true, demand functions are given by the nonlinear model

$$D_i^*(p_1, p_2, \xi_i) = \log(p_i) + \theta_{i1}^* p_1 + \theta_{i2}^* p_2 + \theta_{i3}^* \xi_i + \theta_{i4}^*, \quad i = 1, 2$$

with $\boldsymbol{\theta}_1^* = [-1.2, .5, 1, -9]^T$ and $\boldsymbol{\theta}_2^* = [.3, -1, 1, -9]^T$. We assume (for now) that although we know the parametric form of these demand functions, we do not know the precise values of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and seek to estimate them. The corresponding marginal revenue functions are

$$M_i^*(p_1, p_2, \xi_i; \boldsymbol{\theta}_i^*) = \log(p_i) + \theta_{i1}^* p_1 + \theta_{i2}^* p_2 + \theta_{i3}^* \xi_i + \theta_{i4}^* + 1 + \theta_{ii}^* p_i, \quad i = 1, 2. \tag{32}$$

Here ξ_1, ξ_2 are random variables representing firm-specific knowledge which change over time ("demand shocks") causing prices to shift.

Our idealized assumption is that $\xi_1 = \xi_2 \equiv \xi$, and ξ is common knowledge to both the firms and to the researcher (full-information). In our simulations, we take ξ

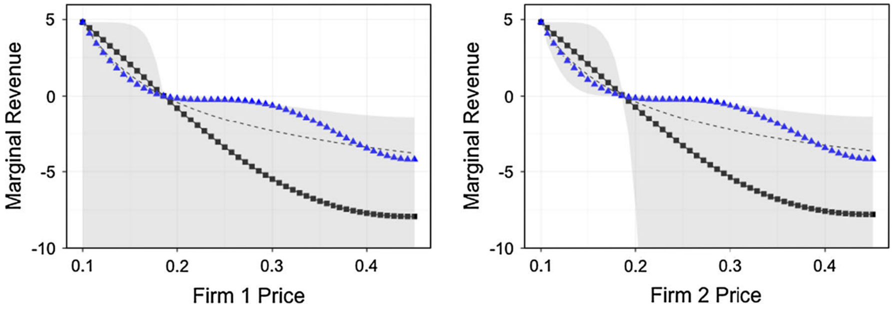


Fig. 1 An idealized scenario. The true marginal revenue function (*dashed black line*), our nonparametric fit (*black line*, *square markers*), and the ambiguity set (*grey region*) for both firms. Every function in the ambiguity set *exactly* reconciles all the data. A sample member (*blue line*, *triangle markers*) shown for comparison. All variables other than the firm's own price have been fixed to the median observation (color figure online)

to be i.i.d normals with mean 5 and standard deviation 1.5. Using these parameters with $\overline{p} = .45$, we simulate values of ξ and solve for the equilibrium prices p_1^j, p_2^j for $j = 1, \ldots, 250$. The values (ξ^j, p_1^j, p_2^j) constitute our data set.

To estimate $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, we substitute the functional form Eq. (32) into Problem (13), adding additional constraints that 1) the marginal revenue of firm i is positive for the minimal price p_i^j observed in the data, 2) the marginal revenue of firm i is decreasing in firm i's price, and 3) a normalization constraint. (See Appendix "Formulation from Sect. 8.1" for an explicit formulation).

Unsurprisingly, solving this optimization recovers the true marginal revenue functions exactly. We say "unsurprisingly" because with full-information a correctly specified, known parametric form, we believe any reasonable estimation procedure should recover the true marginal revenue functions. We point out that the optimal solution to the optimization problem is *unique*, and the optimal value of the residuals is $\boldsymbol{\epsilon} = 0$.

We plot the true marginal revenue functions for each firm (which is the same as our fitted function) in Fig. 1 (dashed black line). To graph these functions we fixed ξ to be its median value over the dataset, and fixed the other firm's price to be the price observed for this median value. For convenience in what follows, we term this type of fixing of the other variables, *fixing to the median observation.*

Next consider the more realistic setting where we do not know the true parametric form (32), and so use our nonparametric method [cf. Problem 22 with dualized objective (25)]. We use a Gaussian kernel and tune the parameter c and regularization constant λ by tenfold cross-validation. The resulting fitted function is shown in Fig. 1 as a black line with square markers. Notice, in particular, this function does not coincide with the true function. However, this function also *exactly* reconciles the data, i.e. the optimal value of the residuals is $\boldsymbol{\epsilon} = 0$. This may seem surprising; the issue is that although there is only one function within the parametric family (32) which reconciles the data, there are many potential smooth, nonparametric functions which also exactly reconcile this data. Using our ambiguity set technique, we compute the upper and lower envelopes of this set of functions, and display the corresponding region as the grey ribbon in Figure 1. We also plot a sample function from this set (blue line with triangle markers).

This multiplicity phenomenon is not unusual; many inverse problems share it. Moreover, it often persists even for very large samples N. In this particular case, the crux of the issue is that, intuitively, the equilibrium conditions only give local information about the revenue function about its minimum. (Notice all three marginal revenue functions cross zero at the same price). The conditions themselves give no information about the global behavior of the function, even as $N \to \infty$.

We see our ambiguity set technique and nonparametric analysis as important tools to protect against potentially faulty inference in these settings. Indeed, parametric estimation might have incorrectly led us to believe that the unique marginal revenue function which recovered the data was the dashed line in Fig. 1—its residual error is zero and it is well-identified within the class. We might then have been tempted to make claims about the slope of the marginal revenue function at the optima, or use it to impute a particular functional form for the demand function. In reality, however, ***any function from the ambiguity set might have just as easily generated this data***, e.g., the blue line with triangle markers. Those previous claims about the slope or demand function, then, need not hold. The data does not support them. Calculating nonparametric sets of plausible alternatives helps guard against these types of unwarranted claims.

Finally, in the absence of any other information, we argue that our proposed nonparametric fit (red line with circles) is a reasonable candidate function in this space of alternatives. By construction it will be smooth and well-behaved. More generally, of all those functions which reconcile the data, it has the smallest $\mathcal{H}$-norm, and thus, by our generalization results in Sect. 7, likely has the strongest generalization properties.

8.2 Bertrand–Nash equilibrium (unobserved effects)

We now proceed to a more realistic example. Specifically, we no longer assume that $\xi_1 = \xi_2$, but rather these values represent (potentially different), firm-specific knowledge that is unobservable to us (the researchers). We assume instead that we only have access to the noisy proxy ξ. In our simulations, we take ξ_1, ξ_2, ξ' to be i.i.d normal with mean 5 and standard deviation 1.5, and let $\xi = (\xi_1 + \xi_2 + \xi')/3$. Moreover, we assume that we we have incorrectly specified that the marginal revenue functions are of the form

$$M_i(p_1, p_2, \xi; \boldsymbol{\theta}_i) = \sum_{k=1}^{9} \theta_{i1k} e^{-kp_1} + \sum_{k=1}^{9} \theta_{i2k} e^{-kp_2} + \theta_{i1} p_1 + \theta_{i2} p_2 + \theta_{i3} \xi_3 + \theta_{i4} \quad (33)$$

for some values of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Notice that the true parametric is not contained in this class. This setup thus includes correlated noise, and endogenous effect, and parametric mispecification. These features are known to cause statistical challenges in simple estimation procedures. We simulate $N = 40$ observations (ξ^j, p_1^j, p_2^j) from this model.

We again fit this model first by solving a modification of Problem (13) as before. (See Appendix "Formulation from Sect. 8.1" for an explicit formulation). We only use

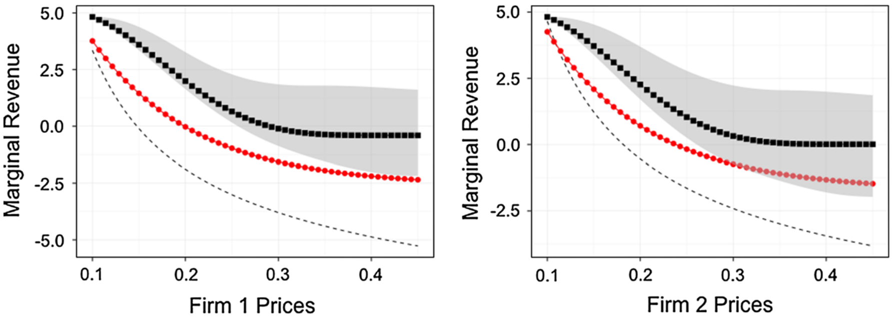


Fig. 2 The true marginal revenue function (*dashed line*), fitted parametric marginal revenue function (*solid red line*, *circular markers*), fitted non-parametric marginal revenue function (*solid black line*, *square markers*)and ambiguity sets (*grey region*) for each firm. We fix all variables except the firm's own price to the median observation (color figure online)

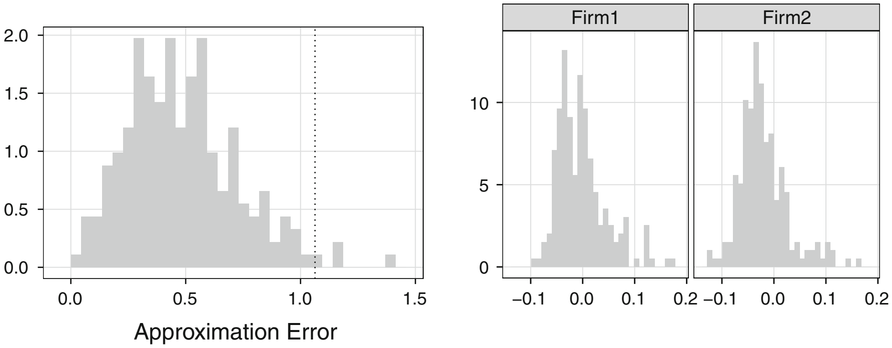


Fig. 3 Bertrand–Nash example of Sect. 8.2. The *left panel* shows the out-of-sample approximation error. The *right panel* shows the out-of-sample prediction error

half the day (20 observations) for reasons that will become clear momentarily. We use the ℓ_∞-norm for the residuals $\boldsymbol{\epsilon}$ and an ℓ_1-regularization of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ in the objective as discussed in Remark 7. We tune the value of λ in the regularization to minimize the mean squared error in price prediction obtaining the value $\lambda = .01$.

Unfortunately, because we used cross-validation to choose λ, Theorem 6 does not directly apply. Consequently, we now refit $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ with $\lambda = .1$ using the other half of our training set. The fitted marginal revenue functions for $\lambda = .01$ can be seen in Fig. 2 (red line, circular markers). Notice that the fitted function does not exactly recover the original function, but does recover its approximate shape.

To assess the out of sample performance of this model, we generate a new set of $N_{out} = 200$ points. For each point we compute the approximation error (minimal ϵ to make this point an ϵ-approximate equilbria), and the prediction error had we attempted to predict this point by the solution to our VI with our fitted function. Histograms of both quantities are in Fig. 3.

The maximal residual on the second half of the training set was $z_N \approx 1.06$, indicated by the dotted line in the left panel. By Theorem 6, we would expect that with at least 90 % probability with respect to the data sampling, a new point would not be an 1.06-

equilibrium with probability at most .21. Our out-of-sample estimate of this probability is .025. In other words, our estimator has much stronger generalization than predicted by our theorem. At the same time, our estimator yields reasonably good predictions. The mean out-of-sample prediction error is $(-.002, 0.02)$ with standard deviation (.048, .047).

Finally, we fit our a nonparametric estimator to this data, using a Gaussian kernel. We again tune the parameter c and regularization constant λ by cross-validation. The resulting fit is shown in Fig. 2 (black line, square markers), along with the corresponding ambiguity set. We chose the value of κ to be twice the standard deviation of the ℓ_1-norm of the residuals, estimated by cross-validation as discussed in Remark 9 and the end of Sect. 7. The out-of-sample approximation error is similar to the parametric case. Unfortunately, the fitted function is not monotone, and, consequently, there exist multiple Nash equilibria. It is thus hard to compare prediction error on the out-of-sample set; which equilibria should we use to predict? This non-monotonicity is a potential weakness of the nonparametric approach in this example.

8.3 Wardrop equilibrium

Our experiments will use the Sioux Falls network [29], a standard benchmark throughout the transportation literature. It is modestly sized with 24 nodes and 76 arcs, and all pairs of nodes represent origin–destination pairs.

We assume that the true function $g(\cdot)$ is given by the US Bureau of Public Roads (BPR) function, $g(t) = 1 + .15t^4$ which is by far the most commonly used for traffic modeling ([14,16]). Baseline demand levels, arc capacities, and free-flow travel times were taken from the repository of traffic problems at [6]. We consider the network structure including arc capacities and free-flow travel times as fixed. We generate data on this network by first randomly perturbing the demand levels a relative amount drawn uniformly from [0, 10 %]. We then use the BPR function to solve for the equilibrium flows on each arc, x_a^*. Finally, we perturb these true flows by a relative amount, again drawn uniformly from [0, 10 %]. We repeat this process $N = 40$ times. Notice that because both errors are computed as relative perturbations, they both are correlated to the observed values. We use the perturbed demands and flows as our data set.

We then fit the function g nonparametrically using (26), again only using half of the data set. The use of low order polynomials in traffic modeling is preferred in the literature for a number of computational reasons. Consequently, we choose k to be a polynomial kernel with degree at most 6, and tune the choice of c by fivefold cross-validation, minimizing the approximation error. The fitted functions for various choices of d are shown in left panel of Fig. 4, alongside the true function. Notice that the fit is quite stable to choice of class of function, and matches the true function very closely. In what remains, we focus on our fit of polynomial of degree 3. We note, this class does not contain the true BPR function (which has degree 4). We refit the degree 3 polynomial with the second-half of our training set (not shown).

To assess the quality of our degree 3 fit, we create a new out-of-sample test-set of size $N_{out} = 500$. On each sample we compute the approximation error of our fit and the ℓ_2-norm of the prediction error when predicting new flows by solving the fitted

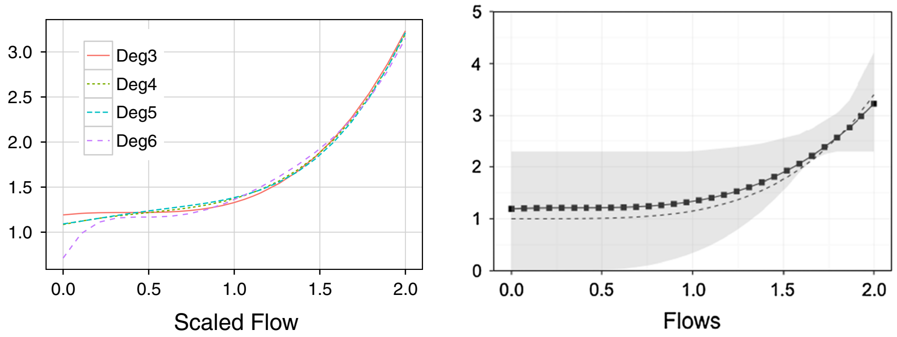


Fig. 4 The *left panel* shows the true BPR function and fits based on polynomials of degrees 3, 4, 5, and 6. The *right panel* shows the true BPR function (*dashed line*), our degree 3 fit (*solid black line* with *markers*), and an ambiguity set around this function (*grey region*) (color figure online)

VI. These numbers are large and somewhat hard to interpret. Consequently we also compute normalized quantities, normalizing the first by the minimal cost of travel on that network with respect to the fitted function and demands, and the second by the ℓ_2 norm of the observed flows. Histograms for the normalized quantities are shown in Fig. 5. The mean (relative) approximation error is 6.5 %, while the mean predictive (relative error) is about 5.5 %.

The in-sample approximation error on the second-half of the training sample was $z_N \approx 8.14 \times 10^5$. By Theorem 7, we can compute that with probability at least 90 % with respect to the data sampling, a new data point will be at most a 9.73×10^5 approximate equilibrium with respect to the fitted function with probability at least 90 %. A cross-validation estimate of the same quantity is 6.24×10^5. Our out of sample estimate of this quantity from the above histogram is 7.78×10^5. In other words, the performance of our estimator is again better than predicted by the theorem. Cross-validation provides a slightly better, albeit biased, bound.

Finally, as in the previous section, we consider constructing an ambiguity set around the fitted function, selecting κ to be two standard deviations as computed by cross-validation. The resulting envelopes are also shown in the right panel of Fig. 4. Notice that in contrast to the envelopes of the previous section, they are quite small, meaning we can have relatively high confidence in the shape of the fitted function.

9 Conclusion

In this paper, we propose a computationally tractable technique for estimation in equilibrium based on an inverse variational inequality formulation. Our approach is generally applicable and focuses on fitting models with good generalization guarantees and predictive power. We prove our estimators enjoy both properties and illustrate their usage in two applications—demand estimation under Nash equilibrium and congestion function estimation under user equilibrium . Our results suggest this technique can successfully model systems presumed to be in equilibrium and make meaningful predictive claims about them.

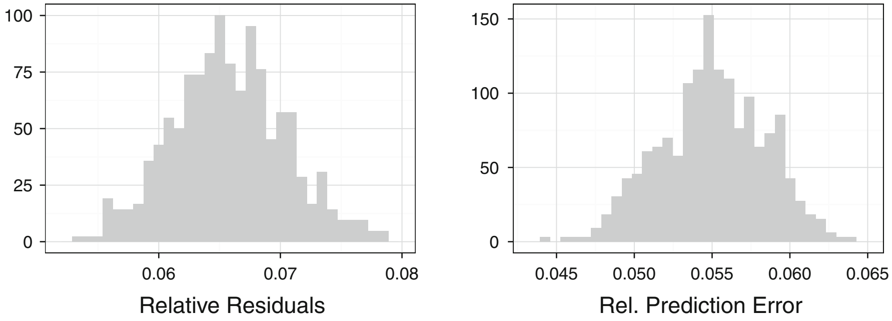


Fig. 5 The *left panel* shows the histogram of of out-sample approximation errors induced by our nonparametric fit from Sect. 8.3. The *right panel* shows the norm of the difference of this flow from the observed flow, relative to the norm of the observed flow

Acknowledgments Research partially supported by the NSF under grant EFRI-0735974, CNS-1239021 and IIS-1237022, by the DOE under grant DE-FG52-06NA27490, by the ARO under grants W911NF-11-1-0227, and W911NF-12-1-0390, by the ONR under grant N00014-10-1-0952, and by Citigroup under a grant to the Sloan School of Management. We would also like to thank the two anonymous reviewers, the Associate editor and the Area editor for their helpful comments on a first draft of this paper.

Appendix 1: Omitted proofs

Proof of Theorem 4

Proof Let $\mathbf{f}^* = (f_1^*, \ldots, f_n^*)$ be any solution. We will construct a new solution with potentially lower cost with the required representation. We do this iteratively beginning with f_1^*.

Consider the subspace $\mathcal{T} \subset \mathcal{H}_1$ defined by $\mathcal{T} = \text{span}(k_{\mathbf{x}_1}, \ldots, k_{\mathbf{x}_N})$, and let $\mathcal{T}^\perp$ be its orthogonal complement. It follows that f_1^* decomposes uniquely into $f_1^* = f_0 + f_0^\perp$ with $f_0 \in \mathcal{T}$ and $f_0^\perp \in \mathcal{T}^\perp$. Consequently,

$$\begin{aligned} f_1^*(\mathbf{x}_j) &= \langle k_{\mathbf{x}_j}, f_1^* \rangle, && \text{[by (20)]} \\ &= \langle k_{\mathbf{x}_j}, f_0 \rangle + \langle k_{\mathbf{x}_j}, f_0^\perp \rangle \\ &= \langle k_{\mathbf{x}_j}, f_0 \rangle && (\text{since } f_0^\perp \in \mathcal{T}^\perp) \\ &= f_0(\mathbf{x}_j) && \text{[by (20)].} \end{aligned}$$

Thus, the solution $\mathbf{f} = (f_0, f_2^*, \ldots, f_n^*)$ is feasible to (22). Furthermore, by orthogonality $\|f_1^*\|_{\mathcal{H}_1} = \|f_0\|_{\mathcal{H}_1} + \|f_0^\perp\|_{\mathcal{H}_1} \geq \|f_0\|_{\mathcal{H}_1}$. Since the objective is non-decreasing in $\|f_1\|_{\mathcal{H}}$, $\mathbf{f}$ has an objective value which is no worse than $\mathbf{f}^*$. We can now proceed iteratively, considering each coordinate in turn. After at most n steps, we have constructed a solution with the required representation. □

Proof of Theorem 5

Proof Suppose Problem (24) is feasible and let $\boldsymbol{\alpha}$ be a feasible solution. Define $\mathbf{f}$ via eq. (23). It is straightforward to check that $\mathbf{f}$ is feasible in Problem (22) with the same objective value.

On the other hand, let $\mathbf{f}$ be some feasible solution to Problem (22). By Theorem 4, there exists $\boldsymbol{\alpha}$ such that $f_i(\mathbf{x}_j) = \mathbf{e}_i^T \boldsymbol{\alpha} \mathbf{K} \mathbf{e}_j$, and $\| f_i \|_{\mathcal{H}}^2 = \mathbf{e}_i^T \boldsymbol{\alpha} \mathbf{K} \boldsymbol{\alpha}^T \mathbf{e}_i$. It straightforward to check that such $\boldsymbol{\alpha}$ is feasible in Problem (24) and that they yield the same objective value. Thus, Problem (22) is feasible if and only if Problem (24) is feasible, and we can construct an optimal solution to Problem (22) from an optimal solution to Problem (24) via (23). □

Proof of Theorem 6

Proof As mentioned in the text, the key idea in the proof is to relate (12) with a randomized uncertain convex program. To this end, notice that if $z_N, \boldsymbol{\theta}_N$ are an optimal solution to (12) with the ℓ_∞-norm, then $(z_N, \boldsymbol{\theta}_N) \in \bigcap_{j=1}^N \mathcal{X}(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ where

$$\mathcal{X}(\mathbf{x}, \mathbf{A}, \mathbf{b}, C) = \left\{ z, \boldsymbol{\theta} \in \Theta : \exists \mathbf{y} \in \mathbb{R}^m \text{ s.t. } \mathbf{A}^T \mathbf{y} \le \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}),\ \ \mathbf{x}^T \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) - \mathbf{b}^T \mathbf{y} \le z \right\}.$$

The sets $\mathcal{X}(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ are convex. Consider then the problem

$$\min_{z \ge 0, \boldsymbol{\theta}} \quad z \quad \text{s.t.} \quad (z, \boldsymbol{\theta}) \in \bigcap_{j=1}^{N} \mathcal{X}(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j).$$

This is exactly of the form Eq. 2.1 in [18]. Applying Theorem 2.4 of that work shows that with probability $\beta(\alpha)$ with respect to the sampling, the "violation probability" of the pair (z_N, θ_N) is a most α. In our context, the probability of violation is exactly the probability that $(\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$ is not a z_N approximate equilibria. This proves the theorem. □

Observe that the original proof in [18] requires that the solution $\boldsymbol{\theta}_N$ be unique almost surely. However, as mentioned on pg. 7 discussion point 5 of that text, it suffices to pick a tie-breaking rule for the $\boldsymbol{\theta}_N$ in the case of multiple solutions. The tie-breaking rule discussed in the main text is one possible example.

Proof of Theorem 7

We require auxiliary results. Our treatment closely follows [7]. Let $\zeta_1, \ldots, \zeta_N$ be i.i.d. For any class of functions $\mathcal{S}$, define the empirical Rademacher complexity $\mathcal{R}_N(\mathcal{S})$ by

$$\mathcal{R}_N(\mathcal{S}) = \mathbb{E}\left[\sup_{f \in \mathcal{S}} \frac{2}{N} \left| \sum_{i=1}^{N} \sigma_i f(\zeta_i) \right| \middle| \zeta_1, \ldots, \zeta_N \right],$$

where σ_i are independent uniform $\{\pm 1\}$-valued random variables. Notice this quantity is random, because it depends on the data $\zeta_1, \ldots, \zeta_N$.

Our interest in Rademacher complexity stems from the following lemma.

Lemma 1 *Let $\mathcal{S}$ be a class of functions whose range is contained in* $[0, M]$. *Then, for any* N, *and any* $0 < \beta < 1$, *with probability at least* $1 - \beta$ *with respect to* $\mathbb{P}$, *every* $f \in \mathcal{F}$ *simultaneously satisfies*

$$\mathbb{E}[f(\zeta)] \leq \frac{1}{N}\sum_{i=1}^{N} f(\zeta_i) + \mathcal{R}_N(\mathcal{S}) + \sqrt{\frac{8M\log(2/\beta)}{N}} \tag{34}$$

Proof The result follows by specializing Theorem 8 of [7]. Namely, using the notation of that work, let $\phi(y, a) = \mathcal{L}(y, a) = a/M$, $\delta = \beta$ and then apply the theorem. Multiply the resulting inequality by M and use Theorem 12, part 3 of the same work to conclude that $M\mathcal{R}_N(M^{-1}\mathcal{S}) = \mathcal{R}_N(\mathcal{S})$ to finish the proof. □

Remark 10 The constants in the above lemma are not tight. Indeed, modifying the proof of Theorem 8 in [7] to exclude the centering of ϕ to $\tilde{\phi}$, one can reduce the constant 8 in the above bound to 2. For simplicity in what follows, we will not be concerned with improvements at constant order.

Remark 11 Lemma 1 relates the empirical expectation of a function to its true expectation. If $f \in \mathcal{S}$ were fixed a priori, stronger statements can be proven more simply by invoking the weak law of large numbers. The importance of Lemma 1 is that it asserts the inequality holds uniformly for all $f \in \mathcal{S}$. This is important since in what follows, we will be identifying the relevant function f by an optimization, and hence it will not be known to us a priori, but will instead depend on the data.

Our goal is to use Lemma 1 to bound the $\mathbb{E}[\epsilon(\mathbf{f}_N, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})]$. To do so, we must compute an upper-bound on the Rademacher complexity of a suitable class of functions. As a preliminary step,

Lemma 2 *For any* $\mathbf{f}$ *which is feasible in* (12) *or* (28), *we have*

$$\tilde{\epsilon}(\mathbf{f}, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C}) \leq \overline{B} \quad a.s. \tag{35}$$

Proof Using strong duality as in Theorem 2,

$$\tilde{\epsilon}(\mathbf{f}, \tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C}) = \max_{\mathbf{x}\in\tilde{\mathcal{F}}} (\tilde{\mathbf{x}} - \mathbf{x})^T \mathbf{f}(\tilde{\mathbf{x}}) \leq 2R \sup_{\tilde{\mathbf{x}}\in\tilde{\mathcal{F}}} \|\mathbf{f}(\tilde{\mathbf{x}})\|_2, \tag{36}$$

by **A6**. For Problem (12), the result follows from the definition of $\overline{B}$. For Problem (28), observe that for any $\tilde{\mathbf{x}} \in \tilde{\mathcal{F}}$,

$$|f_i(\tilde{\mathbf{x}})|^2 = \langle f_i, k_{\tilde{\mathbf{x}}}\rangle^2 \leq \|f_i\|^2_{\mathcal{H}} \sup_{\|\mathbf{x}\|_2 \leq R} k(\mathbf{x}, \mathbf{x}) = \|f_i\|^2_{\mathcal{H}} \overline{K}^2 \leq \kappa_i^2 \overline{K}^2, \tag{37}$$

where the middle inequality follows from Cauchy–Schwartz. Plugging this into Eq. (36) and using the definition of $\overline{B}$ yields the result.

Now consider the class of functions

$$F = \begin{cases} \left\{(\mathbf{x}, \mathbf{A}, \mathbf{b}, C) \mapsto \epsilon(\mathbf{f}, \mathbf{x}, \mathbf{A}, \mathbf{b}, C) : \mathbf{f} = \mathbf{f}(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\right\} & \text{for Problem (12)} \\ \left\{(\mathbf{x}, \mathbf{A}, \mathbf{b}, C) \mapsto \epsilon(\mathbf{f}, \mathbf{x}, \mathbf{A}, \mathbf{b}, C) : f_i \in \mathcal{H}, \;\; \|f_i\|_{\mathcal{H}} \leq \kappa_i \;\; i = 1, \ldots, N\right\} \\ \text{for Problem (28).} \end{cases}$$

Lemma 3

$$\mathcal{R}_N(F) \leq \frac{2\overline{B}}{\sqrt{N}}$$

Proof We prove the lemma for Problem (12). The proof in the other case is identical. Let $\mathcal{S} = \{\mathbf{f}(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. Then,

$$\begin{aligned} \mathcal{R}_N(F) &= \frac{2}{N}\mathbb{E}\left[\sup_{f \in \mathcal{S}} \left|\sum_{j=1}^N \sigma_j \epsilon(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)\right| \middle| (\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)_{j=1}^N\right] \\ &\leq \frac{2\overline{B}}{N}\mathbb{E}\left[\left(\sum_{j=1}^N \sigma_j^2\right)^{\frac{1}{2}}\right] && \text{[using (35)]} \\ &\leq \frac{2\overline{B}}{N}\sqrt{\mathbb{E}\left[\sum_{j=1}^N \sigma_j^2\right]} && \text{(Jensen's inequality)} \\ &= \frac{2\overline{B}}{\sqrt{N}} && (\sigma_j^2 = 1 \text{ a.s.}). \end{aligned}$$

□

We are now in a position to prove the theorem.

Proof (Theorem 7) Observe that $z_N = \frac{1}{N}\sum_{j=1}^N (\epsilon(\mathbf{f}_N, \mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j))^p$. Next, the function $\phi(z) = z^p$ satisfies $\phi(0) = 0$ and is Lipschitz with constant $L_\phi = p\overline{B}^{p-1}$ on the interval $[0, \overline{B}]$. Consequently, from Theorem 12 part 4 of [7],

$$\begin{aligned} \mathcal{R}_N(\phi \circ F) &\leq 2L_\phi \mathcal{R}_N(F) \\ &\leq 2p\overline{B}^{p-1}\frac{2\overline{B}}{\sqrt{N}} \\ &= \frac{4p\overline{B}^p}{\sqrt{N}}. \end{aligned}$$

Now applying Lemma 1 with $\zeta \to (\tilde{\mathbf{x}}, \tilde{\mathbf{A}}, \tilde{\mathbf{b}}, \tilde{C})$, $f(\cdot) \to \epsilon(\cdot)^p$, and $M = \overline{B}^p$ yields the first part of the theorem.

For the second part of the theorem, observe that, conditional on the sample, the event $\tilde{\mathbf{x}}$ is not a $z_N + \alpha$-approximate equilibrium is equivalent to the event that $\epsilon_N > z_N + \alpha$. Now use Markov's inequality and apply the first part of the theorem. □

Proof of Theorem 8

Proof Consider the first part of the theorem.

By construction, $\hat{\mathbf{x}}$ solves VI($\mathbf{f}(\cdot, \theta_N), \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1}$). The theorem, then, claims that $\mathbf{x}_{N+1}$ is $\delta' \equiv \sqrt{\frac{z_N}{\gamma}}$ near a solution to this VI. From Theorem 1, if $\mathbf{x}_{N+1}$ were not δ' near a solution, then it must be the case that $\epsilon(\mathbf{f}(\cdot, \theta_N), \mathbf{x}_{N+1}, \mathbf{A}_{N+1}, \mathbf{b}_{N+1}, C_{N+1})$

$> z_N$. By Theorem 6, this happens only with probability $\beta(\alpha)$.

The second part is similar to the first with Theorem 6 replaced by Theorem 7. □

Appendix 2: Casting structural estimation as an inverse variational inequality

In the spirit of structural estimation, assume there exists a *true* $\boldsymbol{\theta}^* \in \Theta$ that generates solutions $\mathbf{x}_j^*$ to VI($\mathbf{f}(\cdot, \theta^*), \mathbf{A}_j^*, \mathbf{b}_j^*, C_j^*$). We observe $(\mathbf{x}_j, \mathbf{A}_j, \mathbf{b}_j, C_j)$ which are noisy versions of these true parameters. We additionally are given a precise mechanism for the noise, e.g., that

$$\mathbf{x}_j = \mathbf{x}_j^* + \Delta\mathbf{x}_j, \quad \mathbf{A}_j = \mathbf{A}_j^* + \Delta\mathbf{A}_j, \quad \mathbf{b}_j = \mathbf{b}_j^* + \Delta\mathbf{b}_j, \quad C_j = C_j^*,$$

where $(\Delta\mathbf{x}_j, \Delta\mathbf{A}_j, \Delta\mathbf{b}_j)$ are i.i.d. realizations of a random vector $(\tilde{\Delta\mathbf{x}}, \tilde{\Delta\mathbf{A}}, \tilde{\Delta\mathbf{b}})$ and $\tilde{\Delta\mathbf{x}}, \tilde{\Delta\mathbf{A}}, \tilde{\Delta\mathbf{b}}$ are mutually uncorrelated.

We use Theorem 2 to estimate $\boldsymbol{\theta}$ under these assumptions by solving

$$\begin{aligned} \min_{\mathbf{y} \geq \mathbf{0}, \boldsymbol{\theta} \in \Theta, \Delta\mathbf{x}, \Delta\mathbf{A}, \Delta\mathbf{b}} \quad & \left\| \begin{pmatrix} \Delta\tilde{\mathbf{x}}_j \\ \Delta\tilde{\mathbf{A}}_k \\ \Delta\tilde{\mathbf{b}}_j \end{pmatrix}_{j=1,\ldots,N} \right\| \\ \text{s.t.} \quad & (\mathbf{A}_j - \Delta\mathbf{A}_j)^T \mathbf{y}_j \leq_{C_j} \mathbf{f}(\mathbf{x}_j - \Delta\mathbf{x}_j, \boldsymbol{\theta}), \ j = 1, \ldots, N, \\ & (\mathbf{x}_j - \Delta\mathbf{x}_j)^T \mathbf{f}(\mathbf{x}_j - \Delta\mathbf{x}_j, \boldsymbol{\theta}) = \mathbf{b}_j^T \mathbf{y}_j, \ j = 1, \ldots, N, \end{aligned} \qquad (38)$$

where $\| \cdot \|$ refers to some norm. Notice this formulation also supports the case where potentially some of the components of $\mathbf{x}$ are unobserved; simply replace them as optimization variables in the above. In words, this formulation assumes that the "de-noised" data constitute a perfect equilibrium with respect to the fitted $\boldsymbol{\theta}$.

We next claim that if we assume all equilibria occur on the strict interior of the feasible region, Problem (38) is equivalent to a least-squares approximate solution to the equations $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$. Specifically, when $\mathbf{x}^*$ occurs on the interior of $\mathcal{F}$, the VI condition Eq. (1) is equivalent to the equations $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$. At the same time, by Theorem 2, Eq. (1) is equivalent to the system (8, 9) with $\epsilon = 0$ which motivated the constraints in Problem (38). Thus, Problem (38) is equivalent to finding a minimal (with respect to the given norm) perturbation which satisfies the structural equations.

We can relate this weighted least-squares problem to some structural estimation techniques. Indeed, [20] and [37] observed that many structural estimation techniques can be reinterpreted as a constrained optimization problem which minimizes the size of the perturbation necessary to make the observed data satisfy the structural equations, and, additionally, satisfy constraints motivated by orthogonality conditions and the generalized method of moments (GMM). In light of our previous comments, if we augment Problem (38) with the same orthogonality constraints, and all equilibria occur on the strict interior of the feasible region, the solutions to this problem will coincide traditional estimators.

Of course, some structural estimation techniques incorporate even more sophisticated adaptations. They may also pre-process the data (e.g., 2 stage least squares technique in econometrics) incorporate additional constraints (e.g. orthogonality of instruments approach), or tune the choice of norm in the least-squares computation (two-stage GMM estimation). These application-specific adaptations improve the statistical properties of the estimator given certain assumptions about the data generating process. What we would like to stress is that, provided we make the same adaptations to Problem (38)—i.e., preprocess the data, incorporate orthogonality of instruments, and tune the choice of norm—and provided that all equilibria occur on the interior, the solution to Problem (38) must coincide exactly with these techniques. Thus, they necessarily inherit all of the same statistical properties.

Recasting (at least some) structural estimation techniques in our framework facilitates a number of comparisons to our proposed approach based on Problem (12). First, it is clear how our perspective on data alters the formulation. Problem (38) seeks minimal perturbations so that the observed data are exact equilibria with respect to $\boldsymbol{\theta}$, while Problem (12) seeks a $\boldsymbol{\theta}$ that makes the observed data approximate equilibria and minimizes the size of the approximation. Secondly, the complexity of the proposed optimization problems differs greatly. The complexity of Problem (38) depends on the dependence of $\mathbf{f}$ on $\mathbf{x}$ and $\boldsymbol{\theta}$ (as opposed to just $\boldsymbol{\theta}$ for (12)), and there are unavoidable non-convex, bilinear terms like $\Delta \mathbf{A}_j^T \mathbf{y}_j$. These terms are well-known to cause difficulties for numerical solvers. Thus, we expect that solving this optimization to be significantly more difficult than solving Problem (12). Finally, as we will see in the next section, Problem (12) generalizes naturally to a nonparametric setting.

Appendix 3: Omitted formulations

Formulation from Sect. 8.1

Let ξ^{med} be the median value of ξ over the dataset. Breaking ties arbitrarily, ξ^{med} occurs for some observation $j = j^{med}$. Let $p_1^{med}, p_2^{med}, \xi_1^{med}, \xi_2^{med}$ be the corresponding prices and demand shocks at time j^{med}. (Recall that in this section $\xi = \xi_1 = \xi_2$.) These definitiosn make precise what we mean in the main text by "fixing other variables to the median observation. Denote by $\underline{p}_1, \underline{p}_2$ the minimum prices observed over the data set.

Our parametric formation in Sect. 8.1 is

$$\min_{\mathbf{y}, \boldsymbol{\epsilon}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2} \quad \|\boldsymbol{\epsilon}\|_\infty \tag{39a}$$

$$\text{s.t.} \quad \mathbf{y}^j \geq \mathbf{0}, \quad j = 1, \ldots, N,$$

$$y_i^j \geq M_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i), \quad i = 1, 2, \; j = 1, \ldots, N,$$

$$\sum_{i=1}^{2} \overline{p}^j y_i^j - (p_i^j) M_i(p_1^j, p_2^j, \xi^j; \boldsymbol{\theta}_i) \leq \epsilon_j, \quad j = 1, \ldots, N,$$

$$M_1(p_1^j, p_2^{med}, \xi^{med}; \boldsymbol{\theta}_1) \geq M_1(p_1^k, p_2^{med}, \xi^{med}; \boldsymbol{\theta}_1), \;\; \forall 1 \leq j, k \leq N \text{ s.t. } p_1^j \leq p_1^k, \tag{39b}$$

$$M_2(p_1^{med}, p_2^j, \xi^{med}; \boldsymbol{\theta}_2) \geq M_2(p_1^{med}, p_2^k, \xi^{med}; \boldsymbol{\theta}_2), \;\; \forall 1 \leq j, k \leq N \text{ s.t. } p_2^j \leq p_2^k, \tag{39c}$$

$$M_1(\underline{p}_1, p_2^{med}, \xi^{med}; \boldsymbol{\theta}_1) = M_1^*(\underline{p}_1, p_2^{med}, \xi_1^{med}; \boldsymbol{\theta}_1^*) \tag{39d}$$

$$M_2(p_1^{med}, \underline{p}_2, \xi^{med}; \boldsymbol{\theta}_2) = M_2^*(p_1^{med}, \underline{p}_2, \xi_2^{med}; \boldsymbol{\theta}_2^*) \tag{39e}$$

Here M_1 and M_2 are given by Eq. (32). Notice, for this choice, the optimization is a linear optimization problem.

Equations (39b) and (39c) constrain the fitted function to be non-decreasing in the firm's own price. Equations (39d) and (39e) are normalization conditions. We have chosen to normalize the functions to be equal to the true functions at this one point to make the visual comparisons easier. In principle, any suitable normalization can be used.

Our nonparametric formulation is similar to the above, but we replace

- The parametric $M_1(\cdot, \boldsymbol{\theta}_1)$, $M_2(\cdot, \boldsymbol{\theta}_2)$ with nonparametric $M_1(\cdot), M_2(\cdot) \in \mathcal{H}$
- The objective by $\|\boldsymbol{\epsilon}\|_1 + \lambda(\|M_1\|_{\mathcal{H}} + \|M_2\|_{\mathcal{H}})$.

By Theorem 4 and the discussion in Sect. 6, we can rewrite this optimization as a convex quadratic program.

Formulation from Sect. 8.2

Our parametric formulation is nearly identical to the parametric formulation in Appendix "Formulation from Sect. 8.1", with the following changes:

- Replace Eq. (39a) by $\|\boldsymbol{\epsilon}\|_\infty + \lambda(\|\boldsymbol{\theta}_1\|_1 + \|\boldsymbol{\theta}_2\|_1)$
- Replace the definition of M_1, M_2 by Eq. (33).

Our nonparametric formulation is identical to the nonparametric formulation of the previous section.

References

1. Abrahamsson, T.: Estimation of origin–destination matrices using traffic counts—a literature survey. www.iiasa.ac.at/Admin/PUB/Documents/IR-98-021.pdf. (1998)

2. Aghassi, M., Bertsimas, D., Perakis, G.: Solving asymmetric variational inequalities via convex optimization. Oper. Res. Lett. **34**(5), 481–490 (2006). doi:10.1016/j.orl.2005.09.006. http://www.sciencedirect.com/science/article/pii/S0167637705001124
3. Ahuja, R., Orlin, J.: Inverse optimization. Oper. Res. **49**(5), 771–783 (2001)
4. Allon, G., Federgruen, A., Pierson, M.: How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods. Manuf. Serv. Oper. Manag. **13**(4), 489–507 (2011)
5. Bajari, P., Benkard, C., Levin, J.: Estimating dynamic models of imperfect competition. Econometrica **75**(5), 1331–1370 (2007)
6. Bar-Gera, H.: Transportation test problems. http://www.bgu.ac.il/bargera/tntp/ (2011). Accessed Nov 2011
7. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3**, 463–482 (2003)
8. Berry, S., Haile, P.: Nonparametric identification of multinomial choice demand models with heterogeneous consumers. Technical report, National Bureau of Economic Research (2009)
9. Berry, S., Levinsohn, J., Pakes, A.: Automobile prices in market equilibrium. Econom. J. Econom. Soc. **63**(4), 841–890 (1995)
10. Berry, S.T.: Estimating discrete-choice models of product differentiation. RAND J. Econ. **25**(2), 242–262 (1994)
11. Bertsekas, D.: Nonlinear Programming. Athena Scientific, Belmont, Massachusetts (1999)
12. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
13. Braess, D., Nagurney, A., Wakolbinger, T.: On a paradox of traffic planning. Transp. Sci. **39**(4), 446–450 (2005). doi:10.1287/trsc.1050.0127. http://transci.journal.informs.org/content/39/4/446.abstract
14. Branston, D.: Link capacity functions: a review. Transp. Res. **10**(4), 223–236 (1976)
15. Breiman, L., et al.: Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. **16**(3), 199–231 (2001)
16. Bureau of Public Roads: Traffic assignment manual. US Department of Commerce, Urban Planning Division (1964)
17. Campi, M., Carè, A.: Random convex programs with l_1-regularization: sparsity and generalization. SIAM J. Control Optim. **51**(5), 3532–3557 (2013)
18. Campi, M.C., Garatti, S.: The exact feasibility of randomized solutions of uncertain convex programs. SIAM J. Optim. **19**(3), 1211–1230 (2008)
19. Dafermos, S., Nagurney, A.: A network formulation of market equilibrium problems and variational inequalities. Oper. Res. Lett. **3**(5), 247–250 (1984)
20. Dubé, J.P., Fox, J.T., Su, C.L.: Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation. Econometrica **80**(5), 2231–2267 (2012)
21. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Adv. Comput. Math. **13**(1), 1–50 (2000)
22. Fudenberg, D., Tirole, J.: Game Theory. MIT Press, Cambridge (1991)
23. Gallego, G., Huh, W., Kang, W., Phillips, R.: Price competition with the attraction demand model: existence of unique equilibrium and its stability. Manuf. Serv. Oper. Manag. **8**(4), 359–375 (2006)
24. Girosi, F., Jones, M., Poggio, T.: Priors stabilizers and basis functions: from regularization to radial, tensor and additive splines. http://dspace.mit.edu/handle/1721.1/7212 (1993)
25. Guerre, E., Perrigne, I., Vuong, Q.: Optimal nonparametric estimation of first-price auctions. Econometrica **68**(3), 525–574 (2000)
26. Harker, P., Pang, J.: Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. Math. Program. **48**(1), 161–220 (1990)
27. Heuberger, C.: Inverse combinatorial optimization: a survey on problems, methods, and results. J. Comb. Optim. **8**(3), 329–361 (2004)
28. Iyengar, G., Kang, W.: Inverse conic programming with applications. Oper. Res. Lett. **33**(3), 319 (2005)
29. LeBlanc, L., Morlok, E., Pierskalla, W.: An efficient approach to solving the road network equilibrium traffic assignment problem. Transp. Res. **9**(5), 309–318 (1975)
30. Luenberger, D.: Optimization by Vector Space Methods. Wiley-Interscience, New York (1997)
31. Nakayama, S., Connors, R., Watling, D.: A method of estimating parameters on transportation equilibrium models: toward integration analysis on both demand and supply sides. In: Transportation Research Board Annual Meeting 2007 (2007)

32. Nevo, A.: Measuring market power in the ready-to-eat cereal industry. Econometrica **69**(2), 307–342 (2001)
33. Pang, J.: A posteriori error bounds for the linearly-constrained variational inequality problem. Math. Oper. Res. **12**, 474–484 (1987)
34. Perakis, G., Roels, G.: An analytical model for traffic delays and the dynamic user equilibrium problem. Oper. Res. **54**(6), 1151 (2006)
35. Rust, J.: Structural estimation of markov decision processes. Handb. Econom. **4**, 3081–3143 (1994)
36. Smola, A., Schölkopf, B.: Learning with Kernels. MIT press, Cambridge (1998)
37. Su, C.L., Judd, K.L.: Constrained optimization approaches to estimation of structural models. Econometrica **80**(5), 2213–2230 (2012)
38. Trevor, H., Robert, T., Jerome, F.: The elements of statistical learning: data mining, inference and prediction. Springer, New York (2001)
39. Wahba, G.: Spline Models for Observational Data, vol. 59. Society for Industrial Mathematics (1990)
40. Yang, H., Sasaki, T., Iida, Y., Asakura, Y.: Estimation of origin–destination matrices from link traffic counts on congested networks. Transp. Res. B Methodol. **26**(6), 417–434 (1992)
41. Zhao, L., Dafermos, S.: General economic equilibrium and variational inequalities. Oper. Res. Lett. **10**(7), 369–376 (1991)