http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 66, No. 3, May–June 2018, pp. 870–892
ISSN 0030-364X (print), ISSN 1526-5463 (online)

# Inverse Optimization with Noisy Data

**Anil Aswani,**[a] **Zuo-Jun (Max) Shen,**[a,b] **Auyon Siddiq**[c]

[a] Department of Industrial Engineering and Operations Research, University of California, Berkeley, Berkeley, California 94720; [b] Department of Civil and Environmental Engineering, and Tsinghua-Berkeley Shenzhen Institute, University of California, Berkeley, Berkeley, California 94720; [c] Anderson School of Management, University of California, Los Angeles, Los Angeles, California 90095

**Contact:** aaswani@berkeley.edu (AA); maxshen@berkeley.edu, http://orcid.org/0000-0003-4538-8312 (Z-J(M)S); auyon.siddiq@anderson.ucla.edu (AS)

**Received:** July 10, 2015
**Revised:** June 8, 2016; May 1, 2017
**Accepted:** October 5, 2017
**Published Online in Articles in Advance:** May 15, 2018

**Subject Classifications:** statistics: estimation; programming: nonlinear; utility/preference: estimation
**Area of Review:** Optimization

**https://doi.org/10.1287/opre.2017.1705**



**Abstract.** Inverse optimization refers to the inference of unknown parameters of an optimization problem based on knowledge of its optimal solutions. This paper considers inverse optimization in the setting where measurements of the optimal solutions of a convex optimization problem are corrupted by noise. We first provide a formulation for inverse optimization and prove it to be NP-hard. In contrast to existing methods, we show that the parameter estimates produced by our formulation are statistically consistent. Our approach involves combining a new duality-based reformulation for bilevel programs with a regularization scheme that smooths discontinuities in the formulation. Using epi-convergence theory, we show the regularization parameter can be adjusted to approximate the original inverse optimization problem to arbitrary accuracy, which we use to prove our consistency results. Next, we propose two solution algorithms based on our duality-based formulation. The first is an enumeration algorithm that is applicable to settings where the dimensionality of the parameter space is modest, and the second is a semiparametric approach that combines nonparametric statistics with a modified version of our formulation. These numerical algorithms are shown to maintain the statistical consistency of the underlying formulation. Finally, using both synthetic and real data, we demonstrate that our approach performs competitively when compared with existing heuristics.

**Funding:** The authors gratefully acknowledge the support of the National Science Foundation [Award CMMI-1450963] and a Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship. This work was partially supported by the National Science Foundation [Grant CMMI-1265671] and the National Science Foundation of China [Grants 71210002 and 71332005].

**Keywords:** inverse optimization • estimation • statistical learning • semiparametric algorithm

## 1. Introduction

An appreciable share of real-world data represents *decisions*, which can often be characterized as the solutions of correspondingly defined optimization problems. Estimating the parameters of these latent optimization problems has the potential to provide greater insight into how decisions are made, and also enable the prediction of future decisions. Examples of domains where this is important include health systems engineering (Aswani et al. 2016), energy systems engineering (Ratliff et al. 2014), and marketing (Green and Srinivasan 1990), where such estimation may lead to new approaches that enable the individualization of products and incentives. For example, consider a single homeowner who *each day* observes an electricity price and weather forecast and then adjusts the temperature set point for their home's air conditioner. By modeling this homeowner's decision as being generated from an optimization problem, we can directly estimate the price elasticity of comfort—as measured by a standardized function of the temperature set point and the outside temperature (American Society of Heating, Refrigeration, and Air-Conditioning Engineers 2013)—for this particular homeowner. This information is valuable for designing personalized incentive bonus schemes that encourage participation in demand-response programs (Aalami et al. 2010) or promote energy efficiency (Aswani and Tomlin 2012).

### 1.1. Overview

This paper considers the problem of estimating unknown model parameters of an optimization problem based on noisy measurements of its optimal solutions, which is often referred to as *inverse* optimization. In particular, we provide the first *statistical inference* perspective on the inverse optimization problem. This is important because real-world decision data are noisy, either because (i) the data collection process introduces measurement noise, (ii) the decision maker deviates from optimal decisions, phenomena often referred to as *bounded rationality* (Tversky and Kahneman 1981), or (iii) there is mismatch between the parametric form of the model and the true underlying decision-making process.

Noisy data make inverse optimization challenging because noise in the solution data can preclude the

existence of a single set of model parameters that renders all observed solutions exactly optimal. In this setting, the goal of inverse optimization is to find a set of model parameters that achieves a good "fit" with respect to the solution data. More specifically, we are interested in two statistical questions. First, how can we generate estimates of unknown model parameters that asymptotically provide the best possible predictions from the chosen parametric form of the model? In statistics, this property is known as *risk consistency* (Bartlett and Mendelson 2002, Greenshtein and Ritov 2004, Chatterjee 2014). Second, when the chosen model matches the true model that is generating the solution data, how can we generate estimates that asymptotically converge to the true value of the unknown parameters? In statistics, this property is known as *consistency* (Wald 1949, Jennrich 1969, Bickel and Doksum 2006). We will use the term *estimation consistency* to distinguish this concept from risk consistency. Note that estimation consistency generally implies risk consistency.

Restated, a risk consistent estimate asymptotically achieves the lowest possible prediction error (out of all possible predictions permitted by the class of models considered). Hence, risk consistency and estimation consistency allow us to be confident that prediction and estimation accuracy, respectively, will generally improve with additional data. By contrast, an estimator that fails to be risk consistent (so-called *inconsistent* estimators) may yield poor predictions, even if a large amount of data are available. Proving consistency of an estimator is an important topic in the theory of statistical inference (Wald 1949, Jennrich 1969, Bartlett and Mendelson 2002, Greenshtein and Ritov 2004, Bickel and Doksum 2006, Chatterjee 2014, Aswani 2016), and consistency is considered to be a minimal requirement for an estimator (Bickel and Doksum 2006).

The main paper begins with Section 2, which describes the statistical and computational challenges of inverse optimization with noisy data. The section begins by formally defining a (convex) forward optimization problem and its corresponding inverse optimization problem. We specifically formulate the inverse optimization problem such that its solution has the desired statistical consistency properties. Our approach is conceptually similar to least-squares regression in the sense that we also employ a sum-of-squares loss function to fit a parametric model to noisy data. The substantive difference is that inverse optimization involves estimating the (possibly multivalued) solution set of a general convex optimization problem, whereas regression typically involves estimating a (single-valued) function that has a closed-form expression. Because of these differences, much of the classical statistical theory on least-squares regression (Jennrich 1969) is invalid in the inverse optimization setting, and thus new analysis is required. We also note that our approach is not restricted to the use of an $l_2$ norm: results similar to those in our paper can be proved for other loss functions, such as absolute deviation or a likelihood function, but we do not consider those extensions in this paper.

In Section 3, we prove that our inverse optimization formulation produces statistically consistent estimates of the unknown model parameters. The key technical difficulty in proving these results is dealing with continuity issues. In particular, the risk measures are not continuous in the general case but are rather lower semicontinuous. As alluded to above, this precludes the use of the typical statistical machinery used to prove consistency results (namely, the uniform law of large numbers (Jennrich 1969) and related uniform bounds (Bartlett and Mendelson 2002, Greenshtein and Ritov 2004)). To circumvent this difficulty, we define a regularized version of the inverse optimization problem that smooths out any discontinuities, and this regularized version of the problem is constructed using a new duality-based reformulation for bilevel programs. Using epi-convergence theory, we show the regularization parameter can be adjusted to approximate the original inverse optimization problem to arbitrary accuracy. The regularized version of our formulation enables us to prove the desired statistical consistency results.

Section 4 provides two numerical algorithms for solving our formulation of the inverse optimization problem. The first numerical algorithm is an enumeration algorithm that is applicable to settings where the dimensionality of the parameter space is modest (i.e., at most four or five parameters). The second numerical algorithm is a semiparametric approach that combines nonparametric statistics with a modified version of our formulation of the inverse optimization problem. The statistical consistency of these two numerical algorithms is shown using the results from Section 3. Finally, in Section 5 we demonstrate using synthetic and real data sets the competitiveness of our approaches compared with existing heuristics (Keshavarz et al. 2011, Bertsimas et al. 2015).

### 1.2. Literature Review

Existing inverse optimization models differ based on their specification of the *loss function*, and the different models can be broadly categorized into either (i) deterministic settings or (ii) noisy settings. The work in the deterministic setting has primarily focused on single observation situations, wherein a single optimal solution is observed and then used to estimate parameters of the optimization problem. However, in the noisy setting, past work has considered situations with either a single observation or multiple observations.

We begin by describing some of the work in the deterministic setting: Ahuja and Orlin (2001) consider

the estimation of objective function coefficients of general linear programs given a single optimal solution. The feasible region of the inverse problem is formulated using the constraints of the dual program and complemetary slackness conditions. Since the observed solution is assumed to be optimal, feasibility of the inverse problem is guaranteed. Iyengar and Kang (2005) and Zhang and Xu (2010) extend inverse optimization to certain conic forward problems using conic duality theory. Inverse optimization models have also been studied in the context of integer programs (Schaefer 2009, Wang 2009) and network problems (Burton and Toint 1992, Hochbaum 2003, Zhang and Liu 1996). With respect to applications, inverse optimization models have been employed in many different domains, including healthcare (Erkin et al. 2010, Chan et al. 2014), energy (Ratliff et al. 2014, Saez-Gallego et al. 2016), finance (Bertsimas et al. 2012), production planning (Troutt et al. 2006), demand management (Carr and Lovejoy 2000, Bajari et al. 2007), auction design (Beil and Wein 2003), telecommunications (Faragó et al. 2003), and geoscience (Burton and Toint 1992). We refer the reader to Heuberger (2004) for a survey of inverse optimization methods.

The noisy setting has been less studied. Chan et al. (2014) propose a generalized approach to inverse optimization for linear programs where the (single) observed solution may be suboptimal or infeasible. Instead of complementary slackness, the authors use dual feasibility and strong duality to formulate the inverse problem. To accommodate noise, the strong duality constraint is relaxed to guarantee feasibility of the inverse problem. Saez-Gallego et al. (2016) also consider inverse optimization for linear programs and formulate the inverse problem using Karush-Kuhn-Tucker (KKT) conditions. Keshavarz et al. (2011) formulate the inverse problem using the KKT conditions of the optimization problem. To accommodate noise, the KKT conditions are relaxed by introducing slack variables to allow the data to "approximately" satisfy the KKT conditions. Similarly, Bertsimas et al. (2015) consider inverse problems where the observed data are assumed to be in an equilibrium. The authors enforce optimality conditions using a variational inequality and similarly relax the optimality conditions by introducing slack variables to allow the data to "approximately" satisfy the variational inequality.

Our work in this paper is most closely related to the noisy setting with multiple observations that has been previously considered by Keshavarz et al. (2011) and Bertsimas et al. (2015). The key distinction between our work and these two previous approaches is in the choice of the loss function. In Keshavarz et al. (2011) and Bertsimas et al. (2015), the loss function is measured by the amount of slack required to make the measured data satisfy an approximate optimality condition (either the KKT conditions (Keshavarz et al. 2011) or a variational inequality describing optimality (Bertsimas et al. 2015)). By contrast, our approach is to jointly estimate (i) the parameters of the optimization problem and (ii) the denoised versions of the measured data (i.e., the true underlying optimal solutions). By performing this joint estimation, we are able to define our loss function to be the average discrepancy between the measured data and the (estimated) denoised data. As we will show, this difference in loss function leads to significantly improved statistical performance. A secondary distinction is that we propose the use of a novel optimality condition: specifically, we upper bound the objective function of a convex optimization problem by its dual—thereby enforcing a zero duality gap and guaranteeing optimality. An important benefit of using this alternative optimality condition is that it has favorable convexity and continuity properties (which are not available when using KKT conditions or variational inequalities to represent optimality) that enable design of numerical algorithms for solving the inverse optimization problem.

### 1.3. Contributions

Our contributions in this paper include both statistical and optimization results, and there are specifically two main contributions. The first is we show that solving a bilevel formulation for the problem of inverse optimization with noisy data provides parameter estimates that are statistically consistent. This statistical result is independent of the approach used to solve the bilevel formulation. Our second main contribution is to propose two numerical algorithms for solving the bilevel formulation by using a novel duality-based reformulation. However, other numerical algorithms can be used to solve the bilevel formulation. For instance, the bilevel program can be reformulated as a mixed-integer quadratic program (MIQP) in some cases (José Fortuny-Amat 1981, Audet et al. 1997). Our statistical results apply to any numerical algorithm for solving the bilevel formulation, including the MIQP reformulation (when possible) or our two algorithms.

We also prove that existing heuristics for inverse optimization with noisy data (Keshavarz et al. 2011, Bertsimas et al. 2015), which are expressed as convex optimization problems, are statistically inconsistent—meaning that in the limit of increasing amount of data these approaches will generate parameter estimates that converge to incorrect values. This is perhaps not unexpected, because we also prove that the problem of inverse optimization with noisy data is NP-hard. It should be noted that the inverse optimization problem *without* noisy data can be solved in polynomial time, as shown by Keshavarz et al. (2011) and Bertsimas et al. (2015).

An additional contribution is that we propose a novel reformulation of bilevel programs where their lower-level optimization problem is convex. It is common to replace the lower-level problem by the KKT conditions or to upper bound the objective function by the value function (Dempe et al. 2015). However, these approaches face certain numerical difficulties. We propose to upper bound the objective function by its dual, which enforces a zero duality gap and describes an optimal point. The benefit of our optimality condition is it has convexity and continuity properties that support the design of numerical algorithms. The two numerical algorithms we propose directly make use of this optimality condition, and the proofs of our statistical results are also aided by the use of this optimality condition.

### 1.4. Notation

Most notation we use in this paper is standard, and we briefly summarize some of the less usual aspects of our notation. We use $\|\cdot\|$ to denote the usual $l_2$-norm. The indicator function $\mathbb{1}(p)$ is defined to be

$$\mathbb{1}(p) = \begin{cases} 1 & \text{if condition } p \text{ is satisfied,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The notation $[r] = \{1, \ldots, r\}$ refers to sequential set. The Kuratowski limit superior of a sequence of sets $C_\nu \subseteq \mathbb{R}^d$ is defined as

$$\limsup_\nu (C_\nu) = \left\{ x \in \mathbb{R}^d \colon \liminf_\nu \operatorname{dist}(x, C_\nu) = 0 \right\}, \tag{2}$$

where $\operatorname{dist}(x, C) = \inf\{\|x - c\| \mid c \in C\}$. We similarly define $\operatorname{dist}(B, C) = \inf\{\operatorname{dist}(x, C) \mid x \in B\}$.

## 2. Challenges with Noisy Inverse Optimization

This section begins by formalizing the notation for the forward problem, before defining the noisy inverse optimization problem. For the case where we have access to measurements (rather than the underlying distributions), we formulate a related sample average approximation of the inverse optimization problem. We show that both these inverse problems are NP-hard. We conclude by showing that existing heuristic approaches for solving the inverse optimization problem are statistically inconsistent, meaning that in the limit of infinite data these heuristic approaches converge to incorrect solutions.

### 2.1. Model for Forward Problem

Let $x \in \mathbb{R}^d$ be the decision variable, let $u \in \mathbb{R}^m$ be the external input variable, and let $\theta \in \mathbb{R}^p$ be the parameter vector. Then the forward optimization problem is given by

$$\text{FOP} \qquad \min_x \{ f(x, u, \theta) \mid g(x, u, \theta) \leqslant 0 \},$$

where $f \colon \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is a function and $g \colon \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^q$ is a vector-valued function. The solution set of FOP is the set-valued function given by $S(u, \theta) = \arg\min_x \{ f(x, u, \theta) \mid g(x, u, \theta) \leqslant 0 \}$.

The value function of FOP is given by $V(u, \theta) = \min_x \{ f(x, u, \theta) \mid g(x, u, \theta) \leqslant 0 \}$, and the feasible set is defined as $\Phi(u, \theta) = \{ x \in \mathbb{R}^d \colon g(x, u, \theta) \leqslant 0 \}$.

### 2.2. Model for Inverse Optimization Problem

Suppose $(u, y) \in \mathbb{R}^m \times \mathbb{R}^d$ is a vector-valued random variable that is distributed according to some unknown but fixed joint distribution $\mathbb{P}_{(u,y)}$. Let $U \times Y \subseteq \mathbb{R}^m$ be the support of this distribution, meaning the smallest set that satisfies the property $\mathbb{P}_{(u,y)}(U, Y) = 1$. If we define the function

$$\text{RISK} \qquad Q(\theta) = \mathbb{E}\Big( \min_{x \in S(u,\theta)} \|y - x\|^2 \Big),$$

then the inverse optimization problem is given by

$$\text{IOP} \qquad \min\{ Q(\theta) \mid \theta \in \Theta \},$$

where $\Theta \subseteq \mathbb{R}^p$ is a known set. We make the following assumptions.

**Assumption 1** (A1). *The functions $f(x, u, \theta)$ and $g(x, u, \theta)$ are continuous in $x, u, \theta$ and convex in $x$ for fixed $u, \theta$.*

**Assumption 2** (A2). *The set $\Theta$ is convex.*

These assumptions are fairly mild. Assumption A1 is equivalent to stating FOP is a convex optimization problem. Although A2 is necessary for the semiparametric algorithm presented in Section 4 because it ensures polynomial-time computability of the algorithm, it is not necessary for our main results regarding statistical consistency because these results only require that $\Theta$ is well posed. Hence, A2 is one way to ensure $\Theta$ is well posed, and one alternative for which our statistical consistency results would hold is if $\Theta$ is discrete valued and finite.

When the joint distribution $\mathbb{P}_{(u,y)}$ is unknown, we cannot solve IOP without additional information. Fortunately, we can leverage the independent and identically distributed measurements $(u_i, y_i)$ for $i \in [n]$. In principle, we can solve IOP using a sample average approximation:

$$\text{IOP–SAA} \qquad \min\{ Q_n(\theta) \mid \theta \in \Theta \},$$

where

$$\text{RISK–SAA} \qquad Q_n(\theta) = \min_{x_i} \ \frac{1}{n} \sum_{i=1}^n \|y_i - x_i\|^2 \quad \text{s.t. } x_i \in S(u_i, \theta), \quad \forall\, i \in [n].$$

In the context of a decision-making agent, $u_i$ may be interpreted as an external signal the agent responds to and $y_i$ as a noisy observation of the corresponding

decision of the agent. Note that in the expression RISK, the variable $x$ is constrained to be an optimal solution of the forward problem. Similarly, we may interpret $x_i$ as representing an underlying optimal solution (unperturbed by noise) of FOP in the $i$th instance. Note also that while the $u_i$ and $\theta$ are both parameters of FOP, they are different in that the $u_i$ are known and may vary across the $n$ observations, whereas $\theta$ is unknown and is fixed across all instances.

For a concrete example, consider the numerical experiments presented in Section 5.4, where we estimate an individual's utility function capturing the trade-off between maintaing a comfortable indoor temperature versus the amount of energy consumption (and implicitly the air conditioning energy costs) required to cool the room. In that example, the $u$ represents the outside air temperature, $\theta_1$ captures the decision maker's (unknown) trade-off between comfort and energy consumption, $\theta_2$ parameterizes the decision maker's (unknown) preferred temperature (i.e., the preferred temperature is $\theta_2 + u$), $x$ represents the true optimal temperature set point (for the given $u$ and $\theta$), and $y$ represents the temperature set point that we observe.

### 2.3. NP-Hardness of Inverse Optimization Problem

Although all the functions and sets involved in FOP and IOP are convex, solving IOP is NP-hard.

**Theorem 1.** *If* A1 *and* A2 *hold, then* IOP *is NP-hard.*

**Proof.** We prove this by showing a reduction from the problem of computing the best rank 1 approximation of an order 3 tensor (which is NP-hard (Hillar and Lim 2013)) to IOP. Consider any $\psi \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, where $r_1, r_2, r_3 \in \mathbb{Z}_+$. This defines $\psi$ to be an order 3 tensor. We define $\rho = r_1 + r_2 + r_3$, and suppose the parameter vector is given by $\theta = (a, b, c) \in \Theta = \mathbb{R}^\rho$, where $a \in \mathbb{R}^{r_1}$, $b \in \mathbb{R}^{r_2}$, and $c \in \mathbb{R}^{r_3}$. Also define the discrete set $U = [r_1] \times [r_2] \times [r_3]$, where $[r] = \{1, 2, \ldots, r\}$, and suppose that $u = (\alpha, \beta, \gamma)$ is uniformly distributed over $U$. Furthermore, suppose $y$ is a random variable given by $\psi_{\alpha,\beta,\gamma}$, which means that $y$ is dependent on $u$ since $u = (\alpha, \beta, \gamma)$. Then we define the following forward optimization problem:

$$S(u, \theta) = \arg\min_x (x - a_\alpha \cdot b_\beta \cdot c_\gamma)^2. \tag{3}$$

This forward optimization problem is a quadratic program (QP) when $(u, \theta)$ is fixed, and so the solution set is $S(u, \theta) = a_\alpha b_\beta c_\gamma$. Note that the solution set consists of a single point. Next, observe that

$$\min_{\theta \in \Theta} Q(\theta) = \min_{\theta \in \mathbb{R}^\rho} \frac{1}{\rho} \sum_{\alpha=1}^{r_1} \sum_{\beta=1}^{r_2} \sum_{\gamma=1}^{r_3} (\psi_{\alpha,\beta,\gamma} - a_\alpha \cdot b_\beta \cdot c_\gamma)^2, \tag{4}$$

where we have converted the expectation into a weighted sum using the fact that $u$ is uniformly distributed over $U$. Observe that (4) is the problem of computing the best rank 1 approximation to an order 3 tensor (Hillar and Lim 2013). □

**Remark 1.** Inapproximability results for IOP can be shown under the setting where $\Theta$ is allowed to be a discrete set (i.e., A1 holds, but A2 does not hold). In particular, there is a straightforward reduction from the shortest vector problem. This implies that IOP is NP-hard to approximate to within any factor up to $2^{(\log d)^{1-\epsilon}}$, for any $\epsilon \geqslant 0$ (Haviv and Regev 2012).

**Remark 2.** Polynomial-time solvability of IOP is possible in very specific settings. For instance, if FOP is a QP with the solution set $S(u, \theta) = \arg\min_x \{x^2 - 2(\theta + u) \cdot x\} = \theta + u$ or an LP with the solution set $S(u, \theta) = \arg\min_x \{x \colon x = \theta + u\} = \theta + u$, then IOP is a QP: $\min_{\theta \in \Theta} \{\mathbb{E}((y - \theta - u)^2)\}$, and its minimizer is $\theta^* = \mathbb{E}(y - u)$.

In general, since $S(u_i, \theta)$ is the optimal solution sets to FOP under input $u_i$, the problem IOP–SAA is a bilevel program; such programs are usually difficult to solve (Dempe et al. 2015). In fact, IOP–SAA is also NP-hard to solve.

**Remark 3.** In the case where FOP is a linear program, the inverse problem IOP takes the form of a quadratic bilevel program, which is generally NP-hard (Audet et al. 1997). Branch-and-bound algorithms have been proposed for solving such bilevel programs (Bard and Moore 1990).

**Corollary 1.** *If* A1 *and* A2 *hold, then* IOP–SAA *is NP-hard.*

**Proof.** We show this result using the same construction used to prove Theorem 1. In particular, observe that if $\{u_1, \ldots, u_n\} = U$, then IOP–SAA is equivalent to IOP, which is NP-hard by Theorem 1. Finally, note that the condition $\{u_1, \ldots, u_n\} = U$ occurs with nonzero probability since the set $U$ is finite and since the $u_i$ are sampled uniformly from $U$. □

**Remark 4.** Inapproximability results for IOP–SAA can be shown under the setting where $\Theta$ is allowed to be a discrete set (i.e., A1 holds, but A2 does not hold). In particular, the same construction in Remark 1 can be used to show that IOP–SAA is NP-hard to approximate to within any factor up to $2^{(\log d)^{1-\epsilon}}$, for any $\epsilon \geqslant 0$ (Haviv and Regev 2012).

**Remark 5.** Polynomial-time solvability of IOP–SAA is possible in very specific settings. For instance, the constructions in Remark 2 lead to instances of IOP–SAA that are QPs.

### 2.4. Statistical Consistency in Inverse Optimization with Noisy Data

We begin with two statistical definitions of consistency: risk consistency and estimation consistency. These definitions are stated in order of increasing stringency,

meaning that risk consistency is necessary (in situations with sufficient continuity) for estimation consistency. The first definition relates to the best predictions possible using the given forward optimization problem.

**Definition 1** (Risk Consistency). An estimate $\hat{\theta}_n \in \Theta$ is risk consistent if

$$Q(\hat{\theta}_n) \xrightarrow{p} \min\{Q(\theta) \,|\, \theta \in \Theta\}. \tag{5}$$

We should interpret the function $Q(\theta)$ as the expected prediction error when the parameter values are $\theta$, where the prediction is the solution set $S(u,\theta)$. And so the above definition is stating that an estimator $\theta_n$ is risk consistent if the expected prediction error of the estimate $\theta_n$ converges in probability to the minimum prediction error possible when we use the forward optimization model described by FOP and constrain $\theta$ to belong to $\Theta$. In other words, an estimator is risk consistent if it asymptotically provides the best predictions possible.

The second statistical definition relates to the situation where the forward optimization model described by FOP is correct and there is a *true* parameter. In particular, it applies to situations where the below identifiability condition is satisfied. Briefly summarized, the identifiability condition is satisfied when FOP is such that two different parameter values $\theta_1$ and $\theta_2$ lead to two different distributions for measurements of the decision data $y_i$. More details and clarifying examples are found in Appendix B.

**Condition** (Identifiability Condition (IC)). There exists a unique $\theta_0 \in \Theta$ such that the following three subconditions hold: (i) $y = \xi + w$, where $\xi \in S(u,\theta_0)$, $\mathbb{E}(w) = 0$, $\mathbb{E}(w^2) < +\infty$, and $u, \xi$ are independent of $w$; (ii) for all $\theta \in \Theta \backslash \theta_0$, there exists $U(\theta) \subseteq U$ such that $\mathbb{P}(u \in U(\theta)) > 0$ and $\text{dist}(S(u,\theta), S(u,\theta_0)) > 0$ for each $u \in U(\theta)$; and (iii) for each fixed $\theta \in \Theta$, we have $\mathbb{P}(\{u\colon S(u,\theta) \text{ is multivalued}\}) = 0$.

The first subcondition of the identifiability condition is stating that the solution data $y_i$ is a noisy measurement (with noise random variable $w$) of a point that belongs to the solution set $S(u_i,\theta_0)$, and the second subcondition is stating that when $\theta$ is different from $\theta_0$ then this leads to different solution sets. This second subcondition is necessary, because otherwise, we could not distinguish the predictions of FOP when the parameters $\theta$ differ from $\theta_0$. The third subcondition eliminates pathological cases that occur when the solution set at a fixed $\theta$ is so large that it approximately encompasses all possible solutions. Note that this third subcondition is mild, and examples where it is satisfied include when (i) FOP is strictly convex or (ii) FOP is a linear program with random coefficients drawn from a continuous distribution; it holds for other examples as well. The second statistical definition is related to this identifiability condition.

**Definition 2** (Estimation Consistency). Suppose IC holds. An estimate $\hat{\theta}_n \in \Theta$ is estimation consistent if

$$\hat{\theta}_n \xrightarrow{p} \theta_0. \tag{6}$$

Stated in other words, an estimate $\hat{\theta}_n$ is estimation consistent if it converges in probability to the true parameter values $\theta_0$. This is the classical notion of consistency of a statistical estimator (Bickel and Doksum 2006).

Although these statistical notions of consistency are quite natural, it is the case that existing heuristic approaches for solving the inverse optimization problem are statistically inconsistent. We will use VIA to refer to the variational inequality method of Bertsimas et al. (2015), and we refer to the KKT conditions' approach of Keshavarz et al. (2011) as KKA.

**Proposition 1.** *Suppose* A1, A2, *and* IC *hold. Then* VIA (*Bertsimas et al.* 2015) *and* KKA (*Keshavarz et al.* 2011) *are not estimation consistent.*

**Corollary 2.** *Suppose* A1 *and* A2 *hold. Then* VIA (*Bertsimas et al.* 2015) *and* KKA (*Keshavarz et al.* 2011) *are not risk consistent.*

The proofs for Proposition 1 and Corollary 2 are contained in the appendix. The intuition for why VIA and KKA are statistically inconsistent is that they are minimizing an incorrect measure of error: these approaches generate an estimated set of parameters that minimizes the level of suboptimality of the measured solution data. However, this leads to biased estimates because suboptimality is measured by (i) deviations in the value of the objective function of FOP and (ii) the amount of constraint violation of FOP, whereas noise directly perturbs the solution data. This is in contrast to our approach (as exemplified by IOP–SAA) which generates an estimated set of parameters that minimizes the deviation between predicted and measured solution data. This distinction between suboptimality and deviations in the solution data becomes most apparent (and critical) in problems with constraints.

## 3. Consistent Estimation for the Inverse Optimization Problem

Given the statistical inconsistency of existing heuristics, we propose to solve the noisy inverse optimization problem by instead solving SAA–IOP. First, we will need to impose a regularity condition to ensure that FOP and IOP–SAA are numerically well posed.

**Condition** (Regularity Condition (R1)). For each $u \in U$ and $\theta \in \Theta$, the feasible set $\Phi(u,\theta)$ is closed, is bounded, and has a nonempty interior (i.e., $\text{int}(\Phi(u,\theta)) \neq \emptyset$). The feasible set $\Phi(u,\theta)$ is also absolutely bounded, meaning there exists $M > 0$ such that $\|x\| \leqslant M$ for all $x \in \Phi(u,\theta)$, $u \in U$, and $\theta \in \Theta$.

Condition R1 is equivalent to requiring FOP to have a strictly feasible point (i.e., Slater's condition

holds) and that the feasible set of FOP is closed and bounded. The first subcondition requiring the feasible set be closed and bounded is needed to ensure the existence of well-posed primal and dual solutions, and it could be replaced by more general conditions. For instance, we could have instead assumed FOP satisfies the uniform level-boundedness condition (Rockafellar and Wets 1998). We use the above for simplicity of stating the results. The condition that $\Phi(u,\theta)$ has a nonempty interior[1] is needed to ensure continuity of $S(u,\theta)$ through application of the Berge maximum theorem (Berge 1963).

The simplest case of statistical consistency of SAA–IOP occurs when the function $f(x,u,\theta)$ is strictly convex, because of the following result.

**Proposition 2.** *Suppose* A1, A2, *and* R1 *hold. If $f(x,u,\theta)$ is strictly convex in $x$ for fixed $u \in U$ and $\theta \in \Theta$, then $Q_n(\theta)$ is continuous.*

**Proof.** Because the feasible set $\Phi(u,\theta)$ is convex for fixed $u,\theta$ by A1 and has a nonempty interior by R1, this means $\Phi(u,\theta)$ is continuous in $\theta$ by example 5.10 from (Rockafellar and Wets 1998). Thus, we can apply the Berge maximum theorem (Berge 1963) to FOP. This implies that $S(u,\theta)$ is upper hemicontinuous in $\theta$ for fixed $u \in U$. However, $S(u,\theta)$ consists of a single point for fixed $u \in U$ and $\theta \in \Theta$, because the objective function is strictly convex and since R1 holds. Consequently, $S(u,\theta)$ is a continuous single-valued function for fixed $u \in \Theta$ (see, for instance, theorem 2.6 in Rockafellar and Wets 1998). Thus, we can apply the Berge maximum theorem to RISK–SAA, and this implies that $Q_n(\theta)$ as defined in RISK–SAA is continuous. □

In this case, we can prove risk and estimation consistency using standard arguments (Jennrich 1969, van der Vaart 2000, Bickel and Doksum 2006) from statistics that use the uniform law of large numbers (Jennrich 1969). However, this approach cannot be applied to the more general case where $f(x,u,\theta)$ is not strictly convex. In particular, when $f(x,u,\theta)$ is not strictly convex, the function $Q_n(\theta)$ will not generally be continuous. And so a different argument is required because the uniform law of large numbers does not apply to discontinuous functions.

Our approach will be to use a statistical consistency result originally due to Wald (1949) that uses a one-sided bounding argument. The advantage of this approach is that it only requires lower semicontinuity, which we show always holds for $Q_n(\theta)$. However, this result only implies the estimates $\hat{\theta}_n$ converge in probability to the set of minimizers of $Q(\theta)$. This cannot imply risk consistency in the general case because $Q_n(\theta)$ is lower semicontinuous, which means that $Q(\hat{\theta}_n)$ can remain bounded from the minimum $Q(\theta)$. And so for the general case, we will show that a weak risk consistency result holds.

To develop the statistical consistency results for the most general case, we will develop a regularized version of RISK–SAA that is guaranteed to be continuous. The first step of this construction involves proposing a new reformulation for bilevel programs that we call a duality-based reformulation. Next, we use this reformulation to construct a regularized version of RISK–SAA and prove its continuity. We use this regularized version to prove statistical consistency results about IOP–SAA and a regularized version of IOP–SAA.

### 3.1. Duality-Based Reformulation

One approach to solving bilevel problems (such as IOP–SAA) is to reformulate the problem as a normal (i.e., single-level) optimization problem by replacing the constraints $x_i \in S(u_i,\theta)$ with an optimality condition (Dempe et al. 2015). One possibility is to replace $x_i \in S(u_i,\theta)$ by the KKT conditions of FOP, and another possibility is to upper bound the objective function using the value function $f(x_i,u_i,\theta) \leqslant V(u_i,\theta)$. Unfortunately, these approaches often encounter numerical difficulties. The KKT approach leads to a nonlinear program with combinatorial complexity, because of the complimentary slackness in KKT. The value function approach is difficult to implement because closed-form expressions for the value function are not available except for very special cases.

Here, we present a new optimality condition. Given the numerical difficulties of existing approaches, we propose to solve bilevel programs (such as IOP–SAA) by using the Lagrangian dual function to upper bound the objective function. The following proposition shows that our idea of using the dual as an upper bound represents a novel optimality condition.

**Proposition 3.** *Suppose* A1 *and* R1 *hold. Then $x \in S(u,\theta)$ if and only if there exists a corresponding $\lambda \in \mathbb{R}^q$ for which $x$, $\lambda$ satisfy the inequalities*

$$\begin{aligned} f(x,u,\theta) - h(\lambda,u,\theta) &\leqslant 0, \\ g(x,u,\theta) &\leqslant 0, \\ \lambda &\geqslant 0, \end{aligned} \tag{7}$$

*where $h(\lambda,u,\theta)$ is the Lagrangian dual function of* FOP.

Proposition 3 is a consequence of strong duality for convex optimization problems. We can now exactly reformulate RISK–SAA as the following optimization problem:

DB–RISK–SAA

$$\begin{aligned} Q_n(\theta) = \min_{x_i,\lambda_i} \ & \frac{1}{n}\sum_{i=1}^{n} \|y_i - x_i\|^2 \\ \text{s.t.} \ & f(x_i,u_i,\theta) - h(\lambda_i,u_i,\theta) \leqslant 0, \quad \forall\, i \in [n], \\ & g(x_i,u_i,\theta) \leqslant 0, \quad \forall\, i \in [n], \\ & \lambda_i \geqslant 0, \quad \forall\, i \in [n]. \end{aligned}$$

It should be noted that the formulation DB–RISK–SAA requires the Lagrangian dual function $h(\lambda,u,\theta)$ to be

computable in closed form, which is the case for a large class of convex (e.g., linear, quadratic, conic) optimization problems that arise in practice (Boyd and Vandenberghe 2009). In cases where the dual function does not appear to have an analytical representation, we may still solve DB–RISK–SAA by developing an algorithm that computes $h(\lambda, u, \theta)$ numerically, although designing such an algorithm is beyond the scope of this paper.

One important feature of this reformulation is that it is a convex optimization problem for fixed values of $\theta$.

**Proposition 4.** *Suppose* A1 *and* R1 *hold. Then* DB–RISK–SAA *is a convex optimization problem for fixed* $\theta$.

Proposition 4 follows directly from A1 and the concavity of the dual function in $\lambda$.

### 3.2. Regularized Formulation

Recall that $Q_n(\cdot)$ is generally not continuous even when A1, A2, and R1 hold. Consequently, we develop a regularized version of the duality-based problem that is guaranteed to be continuous. We define the $\epsilon$-regularized version of the duality-based problem to be

R–DB–RISK–SAA

$$\begin{aligned} Q_n(\theta;\epsilon) = \min_{x_i,\lambda_i} \ & \frac{1}{n}\sum_{i=1}^{n} \|y_i - x_i\|^2 \\ \text{s.t.} \ & f(x_i,u_i,\theta) - h(\lambda_i,u_i,\theta) \leqslant \epsilon, \quad \forall\, i \in [n], \\ & g(x_i,u_i,\theta) \leqslant \epsilon, \quad \forall\, i \in [n], \\ & \lambda_i \geqslant 0, \quad \forall\, i \in [n]. \end{aligned}$$

We associate this to a regularized version of the sample average approximation of the inverse optimization problem:

R–IOP–SAA $\qquad \min\{Q_n(\theta;\epsilon) \mid \theta \in \Theta\}.$

The idea of this regularization is that we relax the optimality conditions to allow points $x_i$ to be an $\epsilon$-optimal solution. Recall that a point

$$x^\epsilon \in \epsilon\text{-argmin}\{f(x) \mid g(x) \leqslant 0\} \tag{8}$$

if (i) $f(x^\epsilon) - f^* \leqslant \epsilon$ and (ii) $g(x^\epsilon) \leqslant \epsilon$, where $f^* = \min\{f(x) \mid g(x) \leqslant 0\}$.

**Proposition 5.** *Suppose* A1 *and* R1 *hold. Then a point* $x$ *is an* $\epsilon$*-optimal solution if and only if there exists a corresponding* $\lambda \in \mathbb{R}^q$ *for which* $x, \lambda$ *satisfy the inequalities*

$$\begin{aligned} f(x,u,\theta) - h(\lambda,u,\theta) &\leqslant \epsilon, \\ g(x,u,\theta) &\leqslant \epsilon, \\ \lambda &\geqslant 0, \end{aligned} \tag{9}$$

*where* $h(\lambda, u, \theta)$ *is the Lagrangian dual function of* FOP.

One benefit of this regularization is that it ensures convexity of R–DB–RISK–SAA when $\theta$ is fixed.

**Proposition 6.** *Suppose* A1, A2, *and* R1 *hold. Then* R–DB–RISK–SAA *is a convex optimization problem for fixed* $\theta$.

Although the above propositions show that the regularization is equivalent to replacing optimality conditions with $\epsilon$-optimality conditions while maintaining convexity for fixed values of $\theta$, the main benefit of the regularization is that it ensures the function $Q_n(\theta;\epsilon)$ defined in R–DB–RISK–SAA is continuous in $\theta, \epsilon$ for any $\epsilon > 0$.

**Proposition 7.** *Suppose* A1, A2, *and* R1 *hold. Then the function* $Q_n(\theta;\epsilon)$ *is jointly continuous in* $\theta, \epsilon$ *for any* $\epsilon > 0$.

**Proof.** The solution set $S(u,\theta)$ is nonempty under A1 and R1 (see, for instance, Rockafellar and Wets 1998, theorem 1.9). Pick any $x_i \in S(u_i,\theta)$, and let $\lambda_i$ be such that $x_i, \lambda_i$ satisfy (7)—this $\lambda_i$ exists by Proposition 3. Next, consider the sets

$$\begin{aligned} \bar{S}(u_i,\theta;\epsilon) = \{x\colon\ & f(x,u_i,\theta) - h(\lambda_i,u_i,\theta) \leqslant \epsilon, \\ & g(x,u_i,\theta) \leqslant \epsilon\}, \\ S(u_i,\theta;\epsilon) = \{x\colon\ & f(x,u_i,\theta) - h(\lambda,u_i,\theta) \leqslant \epsilon, \\ & g(x,u_i,\theta) \leqslant \epsilon,\ \lambda \geqslant 0\}, \end{aligned} \tag{10}$$

and note that $\bar{S}(u_i,\theta;\epsilon) = S(u_i,\theta;\epsilon)$, since by optimality of $\lambda_i$, with respect to the dual problem, we have $h(\lambda,u_i,\theta) \leqslant h(\lambda_i,u_i,\theta)$ for all $\lambda \geqslant 0$. Observe that the functions $f(x_i,u_i,\theta), g(x_i,u_i,\theta)$ are continuous and convex from A1, and the point $x_i$ belongs to the interior of $\bar{S}(u_i,\theta;\epsilon)$ since it satisfies (7). Thus, we can apply example 5.10 from Rockafellar and Wets (1998). This yields that $\bar{S}(u_i,\theta;\epsilon)$ is continuous in $\theta, \epsilon$ for any $\epsilon > 0$, and so we also get continuity of $S(u_i,\theta;\epsilon)$ by its equality to $\bar{S}(u_i,\theta;\epsilon)$. Since R–DB–RISK–SAA can be written as $Q_n(\theta;\epsilon) = \min_{x_i}\{(1/n)\sum_{i=1}^n \|y_i - x_i\|^2 \mid x_i \in S(u_i,\theta;\epsilon),\ \forall\, i \in [n]\}$, we are able to apply the Berge maximum theorem (Berge 1963). This implies continuity of $Q_n(\theta;\epsilon)$ in $\theta, \epsilon$ for any $\epsilon > 0$. □

A point of note is that within the above proof, we show that the set of $\epsilon$-optimal solutions of a parametric convex optimization problem $S(u_i,\theta;\epsilon)$ is continuous with respect to the parametrization $\theta$; this is in contrast to the solution set of a parametric convex optimization problem $S(u_i,\theta)$, which is in general only upper hemicontinuous with respect to the parametrization $\theta$. The case of a parametric strictly convex optimization problem is the exception, which as shown in the proof of Proposition 2 has a continuous (with respect to the parametrization $\theta$) solution set.

The function $Q_n(\theta;\epsilon)$ will not be jointly continuous in $\theta, \epsilon$ at $\epsilon = 0$. However, it satisfies another property that is useful for solving IOP–SAA.

**Proposition 8.** *Suppose* A1, A2, *and* R1 *hold, and let* $\epsilon_v > 0$ *be a monotone decreasing sequence with* $\epsilon_v \to 0$. *Then*

*we have* $\min\{Q_n(\theta;\epsilon_v) \mid \theta \in \Theta\} \to \min\{Q_n(\theta) \mid \theta \in \Theta\}$ *and*

$$\limsup_v (\arg\min\{Q_n(\theta;\epsilon_v) \mid \theta \in \Theta\}) \subseteq \arg\min\{Q_n(\theta) \mid \theta \in \Theta\}. \quad (11)$$

*If* $z_v > 0$ *is a monotone decreasing sequence with* $z_v \to 0$, *then we also have*

$$\limsup_v (z_v - \arg\min\{Q_n(\theta;\epsilon_v) \mid \theta \in \Theta\}) \subseteq \arg\min\{Q_n(\theta) \mid \theta \in \Theta\}. \quad (12)$$

**Proof.** Let $C_n(\theta,\epsilon)$ be the feasible set of R–DB–RISK–SAA, and define $(X,\Lambda) = \{x_i, \lambda_i, \forall\, i \in [n]\}$. Suppose $(X,\Lambda) \in C_n(\theta,\alpha)$, where $\alpha \geqslant 0$. Then for any $\beta \geqslant \alpha$, we must have $(X,\Lambda) \in C_n(\theta,\beta)$ by the definition of the constraints in R–DB–RISK–SAA. This means that

$$C_n(\theta,\epsilon_1) \supseteq C_n(\theta,\epsilon_2) \supseteq \cdots. \quad (13)$$

As a result, the set $D_n(\theta,\epsilon_v) = \{\theta, X, \Lambda\colon\ \theta \in \Theta$ and $(X,\Lambda) \in C_n(\theta,\epsilon_v)\}$ is also monotone nonincreasing:

$$D_n(\theta,\epsilon_1) \supseteq D_n(\theta,\epsilon_2) \supseteq \cdots. \quad (14)$$

Also, the feasible set $\Phi(u,\theta)$ is convex for fixed $u,\theta$ by A1 and has a nonempty interior by R1. This means $\Phi(u,\theta)$ is continuous in $\theta$ by Rockafellar and Wets (1998, example 5.10), and so we can apply the Berge maximum theorem (Berge 1963) to FOP. This implies that $S(u,\theta)$ is upper hemicontinuous in $\theta$ for fixed $u \in U$. By Dempe et al. (2015, remark 3.2), this means that $Q_n(\theta)$ is lower semicontinuous. Thus, by Rockafellar and Wets (1998, proposition 7.4.d), we have that the extended real-valued function $\{Q_n(\theta;\epsilon_v) \mid \theta \in \Theta\}$ epiconverges to the extended real-valued function $\{Q_n(\theta) \mid \theta \in \Theta\}$. The result then follows from Rockafellar and Wets (1998, exercise 7.32.d and theorem 7.33)). □

**Corollary 3.** *Suppose* A1, A2, *and* R1 *hold. Given any* $d > 0$, *there exist* $E, Z > 0$ *such that if* $\hat{\theta}_n \in z\text{-argmin}\{Q_n(\theta;\epsilon) \mid \theta \in \Theta\}$ *for any* $0 \leqslant z \leqslant Z$ *and* $0 \leqslant \epsilon \leqslant E$, *then* $\operatorname{dist}(\hat{\theta}_n, \arg\min\{Q_n(\theta) \mid \theta \in \Theta\}) < d$.

**Proof.** This is a restatement of Proposition 8. □

These results say that approximately solving R–IOP–SAA is equivalent to approximately solving IOP–SAA.

### 3.3. Statistical Consistency

To prove statistical consistency, we will need to impose an additional regularity condition that ensures expectations of corresponding random variables exist.

**Condition** (Regularity Condition (R2)). The set $\Theta$ is closed and bounded, and $\mathbb{E}(y^2) < +\infty$.

This regularity assumption ensures that the law of large numbers (Wald 1949, Jennrich 1969, van der Vaart 2000) holds in our setting. The above expectation condition holds in many situations, including when $Y$ is bounded or when $y$ has a subexponential distribution (Vershynin 2012). This allows for settings where IC holds with measurement noise that is Gaussian, Bernoulli, bounded support, Laplacian, and exponential, among many other distributions.

Our first statistical consistency result is that solving R–IOP–SAA is risk consistent. To state the result, we must formally define the regularized version of the inverse optimization problem. The regularized risk is

$$\text{R–RISK} \qquad Q(\theta;\epsilon) = \mathbb{E}\Big(\min_{x \in S(u,\theta;\epsilon)} \|y - x\|^2\Big),$$

where $S(u,\theta;\epsilon) = \{x \in \mathbb{R}^d\colon\ f(x,u,\theta) \leqslant V(u,\theta) + \epsilon,\ g(x,u,\theta) \leqslant \epsilon\}$ is the set of $\epsilon$-optimal solutions to FOP. For given $\epsilon > 0$, we define the regularized inverse optimization problem to be

$$\text{R–IOP} \qquad \min\{Q(\theta;\epsilon) \mid \theta \in \Theta\}.$$

The first statistical consistency result specifically concerns nearly optimal solutions of R–IOP–SAA. We say that a sequence of solutions $\hat{\theta}_n$ is nearly optimal for R–IOP–SAA with fixed $\epsilon > 0$ in probability if for any $\delta > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(\operatorname{dist}(\hat{\theta}_n, \arg\min\{Q_n(\theta;\epsilon) \mid \theta \in \Theta\}) > \delta) = 0. \quad (15)$$

**Theorem 2.** *Suppose* A1, A2, R1, *and* R2 *hold. Given any fixed* $\epsilon > 0$, *if* $\hat{\theta}_n$ *is nearly optimal for* R–IOP–SAA *in probability, then we have* $Q(\hat{\theta}_n;\epsilon) \xrightarrow{p} \min\{Q(\theta;\epsilon) \mid \theta \in \Theta\}$.

**Proof.** Proposition 7 gives continuity of $Q_n(\theta;\epsilon)$. Thus, we can apply the uniform law of large numbers (Jennrich 1969), which gives

$$\sup_{\theta\in\Theta} |Q_n(\theta;\epsilon) - Q(\theta;\epsilon)| \xrightarrow{p} 0. \quad (16)$$

Consider any $\theta_0 \in \arg\min\{Q(\theta;\epsilon) \mid \theta \in \Theta\}$ and any $\theta_1 \in \arg\min\{Q_n(\theta;\epsilon) \mid \theta \in \Theta\}$. By assumption $Q_n(\theta_1;\epsilon) \leqslant Q_n(\theta_0;\epsilon)$, and so we have

$$Q(\hat{\theta}_n;\epsilon) + Q_n(\hat{\theta}_n;\epsilon) - Q(\hat{\theta}_n;\epsilon) + Q_n(\theta_1;\epsilon) - Q_n(\hat{\theta}_n;\epsilon) \leqslant Q(\theta_0;\epsilon) + Q_n(\theta_0;\epsilon) - Q(\theta_0;\epsilon). \quad (17)$$

Rearranging terms gives

$$\begin{aligned} Q(\hat{\theta}_n;\epsilon) - Q(\theta_0) \leqslant\ & |Q_n(\hat{\theta}_n;\epsilon) - Q(\hat{\theta}_n;\epsilon)| \\ & + |Q_n(\theta_1;\epsilon) - Q_n(\hat{\theta}_n;\epsilon)| \\ & + |Q_n(\theta_0;\epsilon) - Q(\theta_0;\epsilon)|. \end{aligned} \quad (18)$$

Recall (i) $Q(\theta_0;\epsilon) \leqslant Q(\hat{\theta}_n;\epsilon)$ by definition of $\theta_0$, (ii) $Q_n(\theta;\epsilon)$ is continuous, and (iii) $\hat{\theta}_n$ is nearly optimal for R–IOP–SAA in probability. Thus, combining these facts with (16) and (18) shows that $Q(\hat{\theta}_n;\epsilon) - Q(\theta_0;\epsilon) \xrightarrow{p} 0$. This is the desired result. □

This result says that if we choose any $\epsilon > 0$ and solve R–IOP–SAA to generate an estimate $\hat{\theta}_n$, then the predictions given by the $\epsilon$-optimal solutions to FOP (i.e., $S(u, \hat{\theta}_n; \epsilon)$) are asymptotically the best possible set of predictions when the error of predictions is measured using R–RISK. A stronger risk consistency result is not possible in the general setting because $Q(\theta)$ is typically discontinuous, and so the above result can be interpreted as a weak consistency result.

A stronger risk consistency result is possible in the case where $f(x, u, \theta)$ is strictly convex. We say that a sequence of solutions $\hat{\theta}_n$ is nearly optimal for IOP–SAA in probability if for any $\delta > 0$ we have[2]

$$\lim_{n\to\infty} \mathbb{P}(\text{dist}(\hat{\theta}_n, \arg\min\{Q_n(\theta) \mid \theta \in \Theta\}) > \delta) = 0. \quad (19)$$

**Theorem 3.** *Suppose* A1, A2, R1, *and* R2 *hold. If* $f(x, u, \theta)$ *is strictly convex in* $x$ *(for fixed* $u \in U$ *and* $\theta \in \Theta$*) and* $\hat{\theta}_n$ *is nearly optimal for* IOP–SAA *in probability, then we have* $Q(\hat{\theta}_n) \xrightarrow{p} \min\{Q(\theta) \mid \theta \in \Theta\}$.

**Proof.** Proposition 2 gives continuity of $Q_n(\theta)$. The remainder of the proof is identical to Theorem 2. □

This result says that when FOP is a strictly convex optimization problem and we solve IOP–SAA to generate an estimate $\hat{\theta}_n$, then the predictions given by the solutions to FOP (i.e., $S(u, \hat{\theta}_n)$) are asymptotically the best possible set of predictions when the error of predictions is measured using RISK. The reason it is possible to show risk consistency in this case is that $Q(\theta)$ will be continuous in this setting.

Our final statistical consistency result is that solving IOP–SAA is estimation consistent when IC holds.

**Theorem 4.** *Suppose* A1, A2, R1, R2, *and* IC *hold. If* $\hat{\theta}_n$ *is nearly optimal for* IOP–SAA *in probability, then we have* $\hat{\theta}_n \xrightarrow{p} \theta_0$.

**Proof.** Because the feasible set $\Phi(u, \theta)$ is convex for fixed $u, \theta$ by A1 and has a nonempty interior by R1, this means $\Phi(u, \theta)$ is continuous in $\theta$ by Rockafellar and Wets (1998, example 5.10). Thus, we can apply the Berge maximum theorem (Berge 1963) to FOP. This implies $S(u, \theta)$ is upper hemicontinuous in $\theta$ for fixed $u \in U$. By Dempe et al. (2015, remark 3.2), this means $Q_n(\theta)$ is lower semicontinuous. Thus, we can apply van der Vaart (2000, theorem 5.14).[3] The result follows from the conclusion of van der Vaart (2000, theorem 5.14) if we can show that (i) $\theta_0 \in \arg\min\{Q(\theta) \mid \theta \in \Theta\}$ and (ii) $\theta_0$ is the unique solution. First, note $Q(\theta) = \mathbb{E}(\min_{x \in S(u,\theta)} \|\xi - x\|^2) + \mathbb{E}(w^2)$, since $\xi, x$ is almost surely independent of $w$ because by IC we have that (i) $\xi, u$ are independent of $w$, and (ii) $S(u, \theta)$ is almost surely single valued. Since by IC we have $\xi \in S(u, \theta_0)$, this means that $Q(\theta_0) = \mathbb{E}(w^2)$ and that $\theta_0 \in \arg\min\{Q(\theta) \mid \theta \in \Theta\}$. Next, consider any $\theta \in \Theta \backslash \theta_0$. Then by IC, we have $\mathbb{E}[\min_{x \in S(u,\theta)} \|\xi - x\|^2 \mid u \in U(\theta)] > 0$ since $\xi \in S(u, \theta_0)$ and $\text{dist}(S(u, \theta), S(u, \theta_0)) > 0$ for each $u \in U(\theta)$. Because $\mathbb{P}(u \in U(\theta)) > 0$ from IC, this means $\mathbb{E}(\min_{x \in S(u,\theta)} \|\xi - x\|^2) > 0$ for any $\theta \in \Theta \backslash \theta_0$. Consequently, we have $Q(\theta) > Q(\theta_0)$ for any $\theta \in \Theta \backslash \theta_0$. □

## 4. Numerical Approaches to Solving IOP–SAA

Solving IOP–SAA with $Q_n(\theta)$ as formulated in DB–RISK–SAA is still difficult because it is a nonconvex problem even under A1, A2, and R1. We will propose two approaches to solving this problem. The first is an enumeration algorithm that is applicable to situations where $p$ is modest (i.e., the $\theta \in \mathbb{R}^p$ parameter has between one to five dimensions). The second approach we describe is a semiparametric algorithm, and it can be used in cases where $\theta \in \mathbb{R}^p$ is higher dimensional and the noise term $w$ has a specific distribution. For both algorithms, we will prove that the estimates computed by these methods satisfy the conditions required for statistical consistency.

The difference in the two algorithms is how they trade off computational and statistical performance. The enumeration algorithm requires computation exponential in $p$, while the semiparametric algorithm needs computation polynomial $p$ computation. But the statistical performance of the methods will be the opposite. The estimates and risk of the enumeration algorithm are anticipated to converge at faster rate (with respect to the number of data points) than those of the semiparametric algorithm. The reason is that the semiparametric algorithm makes use of a nonparametric step (via the $l_2$-regularized Nadaraya-Watson estimator, defined in Section 4.2), which is well known to generally converge at a slower rate than a fully parametric approach. Precisely characterizing the statistical convergence rates of the two algorithms is left open for future work.

Although the enumeration algorithm needs exponential in $p$ computation, it is still practical for many real-world problems. Many principal-agent problems (e.g., Zhang and Zenios 2008, Crama et al. 2008) use models where the parameter set is modest in dimensionality (i.e., utility functions with two or three parameters). We demonstrate the practicality of the enumeration algorithm in Section 5 through an energy-related example using real data.

### 4.1. Enumeration Algorithm

The main idea of this algorithm is that computing $Q_n(\theta)$ and $Q_n(\theta; \epsilon)$ for fixed values of $\theta$ can be done in polynomial time since DB–RISK–SAA and R–DB–RISK–SAA are convex optimization problems by Propositions 4 and 6, respectively. This approach enumerates over different fixed values of $\theta$ and solves a series of polynomial-time problems. However, $\Theta$ is a continuous set since because

it is convex by A2. To enable enumeration, we discretize $\Theta$ using a $\delta$-net of $\Theta$, which we will call $\mathcal{T}(\delta)$. (Here, we define this to mean that $\mathcal{T}(\delta)$ is a finite set such that $\max_{\theta\in\Theta}\min_{t\in\mathcal{T}(\delta)}\|t-\theta\| \leqslant \delta$.) We then compute $Q_n(\theta;\epsilon)$ for all $\theta\in\mathcal{T}(\delta)$. And our approximate solution is finally given by $\hat{\theta}_n = \arg\min\{Q_n(\theta;\epsilon)\,|\,\theta\in\mathcal{T}(\delta)\}$.

This approach requires continuity of $Q_n(\theta;\epsilon)$ because otherwise performing an enumeration via the $\delta$-net $\mathcal{T}(\delta)$ may not get sufficiently close to the optimal value. However, $Q_n(\theta;\epsilon)$ is only guaranteed to be continuous at $\epsilon = 0$ when $f(x,u,\theta)$ is strictly convex for fixed $u,\theta$ by Proposition 2 and since $Q_n(\theta;0) = Q_n(\theta)$ by definition. Hence, we require $\epsilon > 0$ for cases where $f(x,u,\theta)$ is *not* strictly convex to ensure continuity of $Q_n(\theta;\epsilon)$ by Proposition 7. Of course, when $f(x,u,\theta)$ is strictly convex, we can set $\epsilon = 0$ and maintain continuity of $Q_n(\theta;\epsilon)$.

This approach is formally presented in Algorithm 1. Importantly, it can be shown that this enumeration algorithm generates nearly optimal solutions of IOP–SAA and R–IOP–SAA. This means the solutions computed by this algorithm satisfy the conditions in Theorems 2–4 that are needed for statistical consistency. In practice, $\epsilon$ is chosen to be $\epsilon = 0$ when FOP is strictly convex, and otherwise $\epsilon$ is chosen to be a small positive value that controls the desired precision of the resulting estimate. An appropriate approach to choose $\epsilon$ and $\delta$ is to use cross-validation, which is a standard data-driven approach from statistics for choosing such parameters (Hastie et al. 2009).

**Algorithm 1** (Enumeration Algorithm)

**Data**: fixed $\delta > 0$ and $\epsilon \geqslant 0$
**Result**: estimate $\hat{\theta}_n$
set $\mathcal{T}(\delta)$ to be $\delta$-net of $\Theta$;
**foreach** $\theta\in\mathcal{T}(\delta)$ **do**
  compute $Q_n(\theta;\epsilon)$ by solving R–DB–RISK–SAA;
set $\hat{\theta}_n \in \arg\min\{Q_n(\theta;\epsilon)\,|\,\theta\in\mathcal{T}(\delta)\}$.

**Theorem 5.** *Suppose* A1, A2, *and* R1 *hold. Given any $d > 0$, there exists $E,\Delta > 0$ such that if $\hat{\theta}_n$ is computed using the enumeration algorithm (i.e., Algorithm 1) for any $0 < \epsilon \leqslant E$ and $0 < \delta \leqslant \Delta$, then* $\operatorname{dist}(\hat{\theta}_n, \arg\min\{Q_n(\theta)\,|\,\theta\in\Theta\}) < d$.

**Proof.** By Corollary 3, there exists $E,Z > 0$ such that if $\hat{\theta}_n \in z\text{-argmin}\{Q_n(\theta;\epsilon)\,|\,\theta\in\Theta\}$ for any $0 \leqslant z \leqslant Z$ and $0 \leqslant \epsilon \leqslant E$, then $\operatorname{dist}(\hat{\theta}_n, \arg\min\{Q_n(\theta)\,|\,\theta\in\Theta\}) < d$. Suppose we choose $z = Z$. Because $Q_n(\theta;\epsilon)$ is continuous in $\theta$ by Proposition 7, there exists $\Delta > 0$ such that for any $0 < \delta \leqslant \Delta$ we have

$$\min\{Q_n(\theta;\epsilon) - Q_n(\theta_0;\epsilon)\,|\,\theta\in\mathcal{T}(\delta)\} < z, \tag{20}$$

where $\theta_0 \in \arg\min\{Q_n(\theta;\epsilon)\,|\,\theta\in\Theta\}$. By construction, we have

$$\arg\min\{Q_n(\theta;\epsilon)\,|\,\theta\in\mathcal{T}(\delta)\} \subseteq z\text{-argmin}\{Q_n(\theta;\epsilon)\,|\,\theta\in\Theta\}. \tag{21}$$

Next, note the enumeration algorithm returns a solution $\hat{\theta}_n \in \arg\min\{Q_n(\theta;\epsilon)\,|\,\theta\in\mathcal{T}(\delta)\}$, which also satisfies $\hat{\theta}_n \in z\text{-argmin}\{Q_n(\theta;\epsilon)\,|\,\theta\in\Theta\}$. The result follows from applying the first line of the proof. □

Theorem 5 states that the estimate obtained using the enumeration algorithm will be at most a distance of $d$ from the set of optimal solutions to IOP-SAA. It immediately follows that for small $d$, the solution of the enumeration algorithm will retain the desirable statistical properties of the solutions to IOP-SAA. As mentioned above, in the special case where FOP is a strictly convex optimization problem, we can simplify the algorithm by setting $\epsilon = 0$. We have a corresponding result about the correctness of the algorithm in this case.

**Theorem 6.** *Suppose* A1, A2, *and* R1 *hold. If $f(x,u,\theta)$ is strictly convex in $x$ (for fixed $u\in U$ and $\theta\in\Theta$), then given any $d > 0$, there exists $\Delta > 0$ such that if $\hat{\theta}_n$ is computed using the enumeration algorithm for $\epsilon = 0$ and any $0 < \delta \leqslant \Delta$, then* $\operatorname{dist}(\hat{\theta}_n, \arg\min\{Q_n(\theta)\,|\,\theta\in\Theta\}) < d$.

**Proof.** By Corollary 3, there exists $E,Z > 0$ such that if $\hat{\theta}_n \in z\text{-argmin}\{Q_n(\theta;\epsilon)\,|\,\theta\in\Theta\}$ for any $0 \leqslant z \leqslant Z$ and $0 \leqslant \epsilon \leqslant E$, then $\operatorname{dist}(\hat{\theta}_n, \arg\min\{Q_n(\theta)\,|\,\theta\in\Theta\}) < d$. Suppose we choose $z = Z$ and $\epsilon = 0$, and note that $Q_n(\theta;0) = Q_n(\theta)$ by their definitions. Because $Q_n(\theta)$ is continuous in $\theta$ by Proposition 2, there exists $\Delta > 0$ such that for any $0 < \delta \leqslant \Delta$ we have

$$\min\{Q_n(\theta) - Q_n(\theta_0)\,|\,\theta\in\mathcal{T}(\delta)\} < z, \tag{22}$$

where $\theta_0 \in \arg\min\{Q_n(\theta)\,|\,\theta\in\Theta\}$. By construction, we have

$$\arg\min\{Q_n(\theta)\,|\,\theta\in\mathcal{T}(\delta)\} \subseteq z\text{-argmin}\{Q_n(\theta)\,|\,\theta\in\Theta\}. \tag{23}$$

Next, note the enumeration algorithm returns a solution $\hat{\theta}_n \in \arg\min\{Q_n(\theta)\,|\,\theta\in\mathcal{T}(\delta)\}$, which also satisfies $\hat{\theta}_n \in z\text{-argmin}\{Q_n(\theta)\,|\,\theta\in\Theta\}$. The result follows from the first line of the proof. □

### 4.2. Semiparametric Approach

Our second approach to solving IOP–SAA is a semiparametric approach. We will need to make an additional assumption about the structure of the problem, as well as impose two more regularity conditions, in order to be able use this approach. We begin with the additional assumption.

**Assumption 3** (A3). *The constraint function $g(x,u,\theta)$ is independent of $\theta$, meaning it can be written as $g(x,u,\theta) = g_0(x,u)$. The objective function $f(x,u,\theta)$ is affine in $\theta$, meaning it can be written as*

$$f(x,u,\theta) = f_0(x,u) + \sum_{j=1}^{p}\theta_j f_j(x,u). \tag{24}$$

Independence of the constraint $g$ from $\theta$ is required because the semiparametric approach relies on fully knowing the feasible region of the forward problem. We note that this is not a particularly strong assumption, since in utility estimation settings one would expect the unknown parameters to appear in the objective function of the forward problem. Keshavarz et al. (2011) and Bertsimas et al. (2015) also assume that the feasible region of the forward problem is independent of the unknown parameters. The second part of A3 ensures that the Lagrangian dual function $h(\lambda, u, \theta)$ is concave in $\theta$. This will enable efficient computation in our semiparametric approach. Next, we describe the two additional regularity conditions.

**Condition** (Regularity Condition (R3)). The objective function $f(x, u, \theta)$ is strictly convex in $x$ (for fixed $u \in U$ and $\theta \in \Theta$) and twice continuously differentiable in $x$, $u$, $\theta$, and the constraints $g(x, u, \theta)$ are continuously differentiable in $x, u, \theta$.

Condition R3 ensures smoothness in the objective function and constraints. The reason we also include a strict convexity assumption is that it acts as a regularity condition: strictly speaking, we require uniqueness of solutions to FOP (which is needed for the denoising step in our semiparametric algorithm) and a second-order growth condition

$$f(x, u, \theta) \geqslant V(u, \theta) + c \cdot [\text{dist}(x, S(u, \theta))]^2 \tag{25}$$

for some $c > 0$ and all $x \in \Phi(u, \theta)$ (which ensures Hölder continuity of the solution set $S(u, \theta)$ with degree $1/2$; Bonnans and Shapiro 2000). Unfortunately, this growth condition can be difficult to directly check even though it has been completely characterized for convex optimization problems (Bonnans and Ioffe 1995a). Fortunately, strict convexity with Slater's constraint qualification (which holds under R1) implies both uniqueness of solutions to FOP and this second-order growth condition (Bonnans and Ioffe 1995b). Hence R3 is sufficient for proving statistical convergence using our algorithm. We also note that our results could be extended to the case where the problem satisfies the first-order growth condition

$$f(x, u, \theta) \geqslant V(u, \theta) + c \cdot \text{dist}(x, S(u, \theta)) \tag{26}$$

for some $c > 0$ and all $x \in \Phi(u, \theta)$. Under this alternative growth condition, the solution set is Hölder continuous with degree 1 (instead of $1/2$). This affects the bound expression in Proposition 9 slightly but otherwise does not qualitatively change our results.

**Condition** (Regularity Condition (R4)). The noise random variable $w$ has a subexponential distribution, meaning that there exists $c > 0$ such that $\mathbb{P}(|w| > t) \leqslant \exp(1 - t/c)$. Also, the probability density function $\mu(u)$ of $u$ is continuously differentiable and is bounded from zero (i.e., $\min_{u \in U} \mu(u) > 0$).

This regularity condition ensures that the distributions of the random variables $w, u$ are not extreme. Most commonly used heavy-tailed noise distributions are subexponential distributions, and so R4 is satisfied by Gaussian, Bernoulli, bounded support, Laplacian, and exponential, among many other distributions (Vershynin 2012). Also, the regularity condition on $\mu(u)$ implies $U$ is bounded.

The idea behind the semiparametric approach is the observation that R–DB–RISK–SAA is convex in $\theta$ for fixed $x$ when A3 holds. However, because the $y_i$ are measured with noise, we cannot simply make the substitution $x_i = y_i$. To overcome this difficulty, we first denoise the $y_i$ using a nonparametric estimator. Specifically, we define the $l_2$-regularized Nadaraya-Watson (L2NW) estimator (Aswani et al. 2013) as

$$\bar{x}_i = \frac{\gamma^{-m} \cdot (1/n) \sum_{j=1}^n y_j \cdot K((u_j - u_i)/\gamma)}{\sigma + \gamma^{-m} \cdot (1/n) \sum_{j=1}^n K((u_j - u_i)/\gamma)}, \tag{27}$$

where $\gamma > 0$ is the *bandwidth* parameter, $\sigma > 0$ is the $l_2$*-regularization* parameter, and $K$: $\mathbb{R}^m \to \mathbb{R}$ is a *kernel function* that satisfies the following properties: (i) $K(u) \geqslant 0$, (ii) $K(u) = 0$ for $\|u\| > 1$, (iii) $K(u) = K(-u)$, and (iv) $\int K(u)\, du = 1$. A common example of a kernel function is the Epanechnikov kernel, which is defined as the function

$$K(u) = \begin{cases} \frac{3}{4} \cdot (1 - \|u\|^2) & \text{if } \|u\| \leqslant 1, \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

The L2NW estimator (27) is computed in polynomial time, and it serves to denoise the $x_i$ in the manner described by the following proposition.

**Proposition 9.** *Suppose* A1 *and* R1–R4 *hold. If* $\gamma = O(n^{-2/(8m+1)})$ *and* $\sigma = O(\gamma)$, *then* $S(u, \theta)$ *consists of a single point, and for sufficiently large* $n$, *we have*

$$\mathbb{P}\left( \max_{i \in [n]} \|\bar{x}_i - S(u_i, \theta_0)\| > n^{-1/(18m)} \right) \leqslant k_1 \exp(-k_2 n^{1/4}), \tag{29}$$

*where* $k_1, k_2 > 0$ *are constants. In particular, this implies* $\max_{i \in [n]} \|\bar{x}_i - S(u_i, \theta_0)\| \xrightarrow{p} 0$.

**Proof.** The first part follows from the strict convexity assumption in R3, and the third part follows directly from the second part. And so we focus on proving the second part. We will prove this using a truncation argument (see, for instance, Tao 2012).

First, note that the function $\psi(x, y) = x/y$ over the domain $(x, y) \in [-M, M] \times [\sigma, \sigma + 1]$ is Lipschitz continuous with constant $L_1 = \sqrt{(M^2 + (\sigma + 1)^2)}/\sigma^2$. Suppose we choose $M = \max_{u \in U} \|\mu(u) S(u, \theta_0)\| + 1$. As a result, using Lemmas 1 and 2, we have

$$\begin{aligned} &\mathbb{P}\left( \left\| \bar{x}_i - \frac{\mu(u_i) S(u_i, \theta_0)}{\sigma + \mu(u_i)} \right\| > t \right) \\ &\quad \leqslant \mathbb{P}\left( \left| \gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n K\left( \frac{u_j - u_i}{\gamma} \right) - \mu(u_i) \right| > \frac{t}{L_1} \right) \end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}\left(\left\|\gamma^{-m}\cdot\frac{1}{n}\sum_{j=1}^{n}y_j\cdot K\left(\frac{u_j-u_i}{\gamma}\right)\right\|>M\right)\\
&+\mathbb{P}\left(\left\|\gamma^{-m}\cdot\frac{1}{n}\sum_{j=1}^{n}y_j\cdot K\left(\frac{u_j-u_i}{\gamma}\right)\right.\right.\\
&\qquad\left.\left.-\mu(u_i)S(u_i,\theta_0)\right\|>\frac{t}{L_1}\right)\\
&\leqslant 2\exp\left(-2c_2n\gamma^{2m}\cdot(t/L_1-c_1\cdot\gamma)^2\right)\\
&\quad+2\exp\left(-2c_2n\gamma^{2m}\cdot(1-c_1\cdot\gamma)^2\right)\\
&\quad+2\exp(-2c_5n\gamma^{2m}\cdot(t/L_1-c_3\cdot\gamma^{1/2}-c_4\cdot\gamma)) \qquad (30)
\end{aligned}
$$

for $t>\max\{c_1\cdot\gamma, c_3\cdot\gamma^{1/2}+c_4\cdot\gamma\}$. Next, observe that the function $\psi(x,y)$ over the domain

$$
\begin{aligned}
(x,y)\in&\left[\min_{u\in U}\mu(u)S(u,\theta),\max_{u\in U}\mu(u)S(u,\theta)\right]\\
&\times\left[\min_{u\in U}\mu(u),\max_{u\in U}\mu(u)\right] \qquad (31)
\end{aligned}
$$

is Lipschitz continuous with some constant $L_2>0$ since (i) the denominator of $\psi$ is bounded away from zero because of R4, and (ii) the numerator of $\psi$ is bounded by R1 and R4. Thus, we have

$$
\begin{aligned}
&\mathbb{P}(\|\bar{x}_i-S(u_i,\theta_0)\|>t)\\
&\quad\leqslant\mathbb{P}\left(\left\|\bar{x}_i-\frac{\mu(u_i)S(u_i,\theta_0)}{\sigma+\mu(u_i)}\right\|>\frac{t-\sigma}{L_2}\right) \qquad (32)
\end{aligned}
$$

for $t>\sigma/L_2$. Suppose we choose $\gamma=O(n^{-2/(8m+1)})$, $\sigma=O(\gamma)$, and $t=n^{-1/(16m+2)}$. Then combining (30) and (32) shows that for sufficiently large $n$, we have

$$
\mathbb{P}(\|\bar{x}_i-S(u_i,\theta_0)\|>n^{-1/(16m+2)})\leqslant c_6\exp(-c_7n^{1/2}), \qquad (33)
$$

where $c_6,c_7>0$ are constants. And so combining the union bound with (33) gives

$$
\begin{aligned}
&\mathbb{P}\left(\max_{i\in[n]}\|\bar{x}_i-S(u_i,\theta_0)\|>n^{-1/(16m+2)}\right)\\
&\quad\leqslant n\mathbb{P}(\|\bar{x}_i-S(u_i,\theta_0)\|>n^{-1/(16m+2)})\\
&\quad\leqslant c_6\exp(-c_7n^{1/2}+\log n). \qquad (34)
\end{aligned}
$$

The final implication of the result follows by noting that $n^{-2/(8m+1)}\to 0$ and $c_6\exp(-c_7n^{1/2}+\log n)\to 0$ as $n\to\infty$. □

Before we present our algorithm, we need one more result that provides additional understanding for the semiparametric approach. Consider the following optimization problem:

ROBUST–IOP–SAA $\qquad \min_{\theta}\max_{\epsilon\geqslant 0}\{Q_n(\theta;\epsilon)\mid\theta\in\Theta\}.$

**Proposition 10.** *Suppose* A1, A2, *and* R1 *hold. Then the solution sets in* $\theta$ *of* ROBUST–IOP–SAA *and* IOP–SAA *are equivalent, and the optimal value of* ROBUST–IOP–SAA *occurs at* $\epsilon=0$.

**Proof.** Let $C_n(\theta,\epsilon)$ be the feasible set of R–DB–RISK–SAA. As shown in the proof for Proposition 8, the feasible set satisfies

$$
C_n(\theta,0)\subseteq C_n(\theta,\epsilon) \qquad (35)
$$

for all $\epsilon\geqslant 0$. As a result, we must have that $Q_n(\theta;0)\geqslant Q_n(\theta;\epsilon)$ for all $\epsilon\geqslant 0$. This means that $\max_{\epsilon\geqslant 0}Q_n(\theta;\epsilon)=Q_n(\theta;0)$. The result holds because $Q_n(\theta;0)=Q_n(\theta)$, by definition. □

Given the above relationship that the optimal value of ROBUST–IOP–SAA occurs at $\epsilon=0$, we propose to solve the inverse optimization problem using the following formulation:

SP–IOP–RISK–SAA

$$
\begin{aligned}
\hat{\theta}_n\in\arg\min\ &\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\\
\text{s.t. }&f(\bar{x}_i,u_i,\theta)-h(\lambda_i,u_i,\theta)\leqslant\epsilon_i,\quad\forall\, i\in[n],\\
&\lambda_i\geqslant 0,\quad\forall\, i\in[n],
\end{aligned}
$$

where the $\bar{x}_i$ are as defined in (27). This is a convex optimization problem.

**Proposition 11.** *Suppose* A1–A3 *and* R1 *hold. Then* SP–IOP–RISK–SAA *is a convex optimization problem.*

We now have the elements to construct our semiparametric algorithm, which is a two-step approach. In the first step, we denoise the $y_i$ data using the L2NW estimator given in (27). This denoising step produces an estimate of the true underlying optimal solution, which we represent by $\bar{x}_i$. While the estimates $\bar{x}_i$ are asymptotically (in $n$) optimal (Proposition 9), they may be suboptimal at finite $n$. Therefore, in the second step, we solve SP–IOP–RISK–SAA, which produces a parameter estimate $\hat{\theta}_n$ that minimizes the suboptimality of $\bar{x}_i$. This approach maintains statistical consistency because the $\bar{x}_i$ are denoised, and it is formally presented in Algorithm 2. Importantly, it can be shown that this semiparametric algorithm generates nearly optimal solutions of IOP–SAA. This means the solutions computed by this algorithm satisfy the conditions in Theorems 2–4 that are needed for statistical consistency. In practice, the values of $\sigma$ and $\gamma$ can be chosen using cross-validation, which is a standard data-driven approach from statistics for choosing such parameters (Hastie et al. 2009).

**Algorithm 2** (Semiparametric Algorithm)

**Data**: fixed $\gamma>0$ and $\sigma>0$
**Result**: estimate $\hat{\theta}_n$
**foreach** $i\in[n]$ **do**
  compute $\bar{x}_i$ using using (27);
compute $\hat{\theta}_n$ using SP–IOP–RISK–SAA;

**Theorem 7.** *Suppose* A1–A3 *and* R1–R4 *and* IC *hold. If* $\sigma = O(n^{-2/(8m+1)})$, $\lambda = O(\sigma)$, *and* $\hat{\theta}_n$ *are computed using the semiparametric algorithm (i.e., Algorithm 2), then* $\hat{\theta}_n$ *is nearly optimal for* IOP–SAA *in probability.*

**Proof.** Note that we have $\min\{-h(\lambda,u,\theta) \,|\, \lambda \geqslant 0\} = -f(S(u,\theta),u,\theta)$ by strong duality (which holds because of A1 and R1; see Bonnans and Shapiro 2000). Next, consider the function

$$\begin{aligned} R(\theta) &= \mathbb{E}\Big(\min_{\lambda \geqslant 0} f(S(u,\theta_0),u,\theta) - h(\lambda,u,\theta)\Big) \\ &= \mathbb{E}(f(S(u,\theta_0),u,\theta) - f(S(u,\theta),u,\theta)), \end{aligned} \tag{36}$$

its sample average approximation

$$\begin{aligned} R_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\min_{\lambda_i \geqslant 0} f(S(u_i,\theta_0),u_i,\theta) - h(\lambda_i,u_i,\theta)\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}(f(S(u_i,\theta_0),u_i,\theta) - f(S(u_i,\theta),u_i,\theta)), \end{aligned} \tag{37}$$

and its semiparametric approximation

$$\begin{aligned} \bar{R}_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\min_{\lambda_i \geqslant 0} f(\bar{x}_i,u_i,\theta) - h(\lambda_i,u_i,\theta)\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}(f(\bar{x}_i,u_i,\theta) - f(S(u_i,\theta),u_i,\theta)). \end{aligned} \tag{38}$$

Note that $\min\{\bar{R}_n(\theta) \,|\, \theta \in \Theta\}$ is simply a reformulation of the problem SP–IOP–RISK–SAA. Next, observe that $\mathbb{E}[f(S(u,\theta_0),u,\theta) - f(S(u,\theta),u,\theta) \,|\, u \in U(\theta)] > 0$ since (i) $f(x,u,\theta)$ is twice continuously differentiable in $x$ by R3, and (ii) $\text{dist}(S(u,\theta),S(u,\theta_0)) > 0$ for each $u \in U(\theta)$ by IC. Consequently, we have $R(\theta) > 0$ for $\theta \in \Theta \backslash \theta_0$. As shown in the proof for Proposition 2, $S(u,\theta)$ is continuous in $\theta$. And so $R_n(\theta)$ and $\bar{R}_n(\theta)$ are continuous because $f(x,u,\theta)$ is twice continuously differentiable in $x,\theta$ by R3.

Next, recall that $U$ is bounded by R4, $\Theta$ is bounded by R2, $f(x,u,\theta)$ is twice continuously differentiable in $x$, $\theta$ by R3, and the feasible set of FOP is absolutely bounded by R1. This means that there exists $L > 0$ such that for all $\theta \in \Theta$, we have $\max_{i\in[n]} |f(\bar{x},u_i,\theta) - f(S(u_i,\theta_0),u_i,\theta)| \leqslant L n^{-1/(18m)}$ whenever $\max_{i\in[n]} \|\bar{x}_i - S(u_i,\theta_0)\| \leqslant n^{-1/(18m)}$ (which occurs with probability at least $1 - k_1\exp(-k_2 n^{1/4})$ by Proposition 9). Thus, we have that $\sup_{\theta\in\Theta} |R_n(\theta) - \bar{R}_n(\theta)| \xrightarrow{p} 0$. Now consider any $\hat{\theta}_n \in \arg\min\{\bar{R}_n(\theta) \,|\, \theta \in \Theta\}$, and note that the estimate $\hat{\theta}_n$ returned by the semiparametric algorithm satisfies this property by construction. By definition, we have $\bar{R}_n(\hat{\theta}_n) \leqslant \bar{R}_n(\theta_0)$, which can be rewritten as

$$R_n(\hat{\theta}_n) + \bar{R}_n(\hat{\theta}_n) - R_n(\hat{\theta}_n) \leqslant R_n(\theta_0) + \bar{R}_n(\theta_0) - R_n(\theta_0). \tag{39}$$

Thus, we have

$$R_n(\hat{\theta}_n) \leqslant R_n(\theta_0) + |\bar{R}_n(\hat{\theta}_n) - R_n(\hat{\theta}_n)| + |\bar{R}_n(\theta_0) - R_n(\theta_0)|. \tag{40}$$

We have thus shown all the conditions required to apply van der Vaart (2000, theorem 5.14), which gives $\hat{\theta}_n \xrightarrow{p} \theta_0$. Now let $\bar{\theta}_n \in \arg\min\{Q_n(\theta) \,|\, \theta \in \Theta\}$. By Theorem 4, we have $\bar{\theta}_n \xrightarrow{p} \theta_0$. This means that $|\bar{\theta}_n - \hat{\theta}_n| \xrightarrow{p} 0$. □

Theorem 7 states that the semiparametric algorithm produces estimates that are statistically consistent under the appropriate conditions. In the next section, we present several numerical experiments that validate our theoretical results as well as the performance of the enumeration and semiparametric algorithms.

## 5. Numerical Experiments

We present numerical results that demonstrate the statistical consistency of our algorithms for inverse optimization with noisy data, and the results show that our algorithms perform competitively against KKA (Keshavarz et al. 2011) and VIA (Bertsimas et al. 2015). We begin by conducting two types of tests using synthetic data. The first type is where the model is kept fixed and the number of data points increases, and the purpose is to demonstrate either estimation consistency or risk consistency of our algorithms. The second type is where the number of data points is kept fixed and the number of the parameters in the model increases, and the purpose is to demonstrate the feasibility of using our algorithms on large-scale problems. We then apply our framework to a real data set, where we estimate a utility function that describes the trade-off made between occupant comfort and energy consumption when setting a thermostat temperature set point for air conditioning.

### 5.1. Synthetic Data and Enumeration Algorithm

In the first experiments, we generate data using a given FOP and then use the same set of equations in SAA-IOP. In other words, the first set of experiments are situations where the model whose parameters are being identified exactly match the model that generates the data. As a result, this setting consists of situations where IC is satisfied. The first example is where (i) FOP-A is $\min\{(\theta + u)\cdot x \,|\, x \in [-1,1]\}$, (ii) $u$ has a uniform distribution with support $[-1,1]$, (iii) the measurement noise $w$ has a normal distribution with zero mean and unit variance, (iv) the data are generated with $\theta_0 = 1$, and (v) the enumeration algorithm (i.e., Algorithm 1) was applied with $\epsilon = 0.001$, $\delta = 0.01$, and $\Theta = [-1,1]$. The second example is where (i) FOP-B is $\min\{x^2 - (\theta + u)\cdot x \,|\, x \in [0,1]\}$, (ii) $u$ has a uniform distribution with support $[0,2]$, (iii) the measurement noise $w$ has a normal distribution with zero mean and unit variance, (iv) the data are generated with $\theta_0 = \frac{1}{2}$,

**Table 1.** Estimation Error $|\hat{\theta}_n - \theta_0|$ of Enumeration Algorithm (ENA) and Benchmark Algorithms (KKA and VIA) on Two Synthetic Instances ($n$ Increasing, $p = 1$)

| | $n$ | 10 | 30 | 50 | 100 | 300 | 500 | 1,000 |
|---|---|---|---|---|---|---|---|---|
| Data: FOP-A | ENA | 0.2616 | 0.0926 | 0.0380 | 0.0211 | 0.0055 | 0.0030 | 0.0009 |
| Model: FOP-A | KKA | 0.8686 | 0.8293 | 0.8182 | 0.8257 | 0.8130 | 0.8231 | 0.8170 |
| | VIA | 0.5552 | 0.4976 | 0.4829 | 0.4887 | 0.4807 | 0.4846 | 0.4780 |
| Data: FOP-B | ENA | 0.4577 | 0.2481 | 0.1510 | 0.0501 | 0.0222 | 0.0123 | 0.0063 |
| Model: FOP-B | KKA | 0.5065 | 0.2281 | 0.1595 | 0.0751 | 0.0398 | 0.0342 | 0.0238 |
| | VIA | 0.9488 | 0.7051 | 0.6344 | 0.4284 | 0.3145 | 0.3810 | 0.2962 |

and (v) the enumeration algorithm (i.e., Algorithm 1) was applied with $\epsilon = 0$, $\delta = 0.01$, and $\Theta = [0, 2]$.

The results averaged over 100 repetitions of sampling $n \in \{10, 30, 50, 100, 300, 500, 1{,}000\}$ data points and then estimating the parameter $\theta$ are summarized in Table 1. We label the enumeration algorithm (i.e., Algorithm 1) as ENA in the table. These results display estimation consistency of the enumeration algorithm since estimation error is decreasing to zero. To further illustrate estimation consistency, we conducted an experiment with the two examples above where the data were generated with a $\theta_0$ that was randomly chosen from a uniform distribution with support $[-1, 1]$ and $[0, 2]$ for the first and second examples, respectively. A plot comparing the estimates $\hat{\theta}_n$ to the true parameter $\theta_0$ for the first situation when $n = 1{,}000$ is shown in Figure 1, and a plot comparing the estimates $\hat{\theta}_n$ to the true parameter $\theta_0$ for the second situation when $n = 10{,}000$ is shown in Figure 2. Consistent estimates should line up along the diagonal, and hence these plots demonstrate the estimation consistency (inconsistency) of the enumeration algorithm (KKA and VIA). Recall from the discussion in Section 2 that KKA and VIA are inconsistent because they minimize an incorrect measure of error, and this discrepancy is most significant for points where the optimal solution of FOP lies on the boundary of the feasible set. KKA and VIA perform more poorly for FOP-A than for FOP-B because FOP-A is a linear program, which has almost all of its optimal solutions on the boundary of the feasible set, whereas FOP-B is a quadratic program, which has more optimal solutions within the strict interior of the feasible set.

In the second set of experiments, we generate data using a given model that is different than the FOP used

**Figure 1.** (Color online) Scatter Plot Comparing Estimated Parameter $\hat{\theta}_n$ vs. True Parameter $\theta_0$ as Computed by ENA, KKA, and VIA Algorithms at $n = 1{,}000$, When the Data and Model Are Both FOP-A

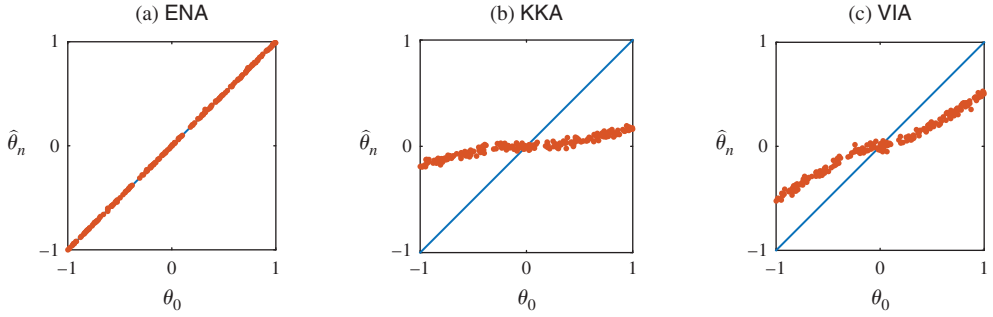
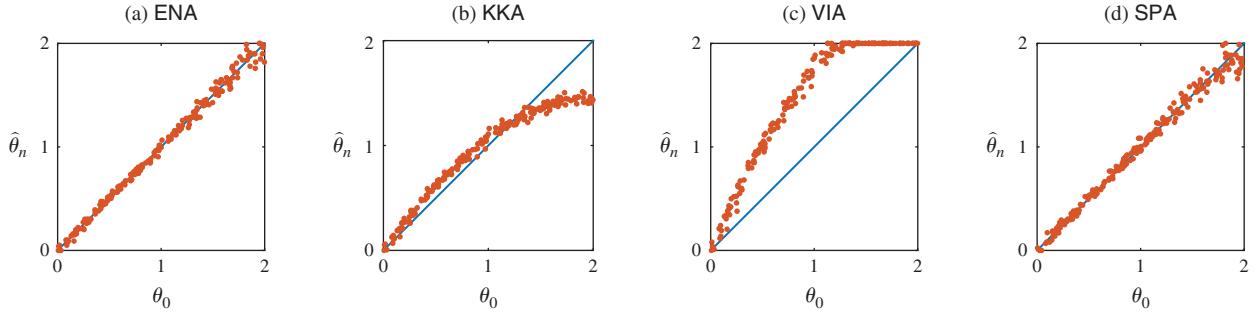



**Figure 2.** (Color online) Scatter Plot Comparing Estimated Parameter $\hat{\theta}_n$ vs. True Parameter $\theta_0$ as Computed by ENA, KKA, VIA, and SPA Algorithms at $n = 10{,}000$ When the Data and Model Are Both FOP-B



to formulate SAA-IOP. In other words, this set of experiments are situations where the model whose parameters are being identified does not match the model that generates the data. As a result, this setting consists of situations where IC is *not* satisfied. The first example is where (i) the data are generated by FOP-C, which is $\min\{\frac{3}{2}\cdot x^2-(1+u)\cdot x \mid x\in[0,1]\}$; (ii) the model estimated by IOP-SAA is FOP-B; (iii) $u$ has a uniform distribution with support $[0,5]$; (iv) the measurement noise $w$ has a normal distribution with zero mean and unit variance; and (v) the enumeration algorithm (i.e., Algorithm 1) was applied with $\epsilon=0$, $\delta=0.01$, and $\Theta=[0,2]$. The second example is where (i) the data are generated by the statistical model SQR-1 given by $y_i=\min\{\max\{\sqrt{u_i},0\},1\}+w_i$; (ii) the model estimated by IOP-SAA is FOP-B; (iii) $u$ has a uniform distribution with support $[0,5]$; (iv) the measurement noise $w$ has a normal distribution with zero mean and unit variance; and (v) the enumeration algorithm (i.e., Algorithm 1) was applied with $\epsilon=0$, $\delta=0.01$, and $\Theta=[0,2]$.

The results averaged over 100 repetitions of sampling $n\in\{10,30,50,100,300,500,1{,}000\}$ data points and then estimating the parameter $\theta$ are summarized in Table 2, and these results are normalized by subtracting $\mathrm{var}(w)$. The reason for this normalization is that the prediction error $\mathbb{E}((y-\xi(u))^2)$ of the prediction $\xi(u)$ of the true model (either FOP-C or SQR-M) is $\mathrm{var}(w)$ because $y=\xi(u)+w$ here. The enumeration algorithm has a lower prediction error because it is risk consistent, whereas KKA and VIA are not risk consistent.

### 5.2. Synthetic Data and Semiparametric Algorithm

We now examine the performance of the semiparametric algorithm (Algorithm 2) in four sets of experiments. In the first set of experiments, we generate data using a given FOP and then use the same equations in SAA-IOP. These experiments are situations where the model whose parameters are being identified exactly matches the model that generates the data. As a result, this setting consists of situations where IC is satisfied. We consider three different formulations for FOP. The first example is where (i) FOP-D is $\min\{x'x-(\theta+u)'x \mid x\in[0,1]^p\}$, (ii) $u$ has a uniform distribution with support $[0,2]^p$, (iii) the measurement noise $w$ has a jointly Gaussian distribution with zero mean and identity covariance, (iv) the data are generated with $p=10$ and $\theta_0\in\mathbb{R}^p$ such that $\theta_{0k}=\frac{1}{2}$ for all $k\in[p]$, and (v) the semiparametric algorithm (i.e., Algorithm 2) was applied with $\gamma,\sigma$ chosen using cross-validation (Hastie et al. 2009) and $\Theta=[0,2]$. The second example is where (i) FOP-E is

$$\min\left\{-\sum_{k=1}^{p}\theta_k\cdot\log(x_k+u_k)-\log(x_{p+1}+u_{p+1})\,\middle|\,x_k\geqslant 0,\ \sum_{k=1}^{p+1}x_k=1\right\}; \quad (41)$$

(ii) $u$ has a uniform distribution with support $[1,2]^{p+1}$; (iii) the measurement noise $w$ has a jointly Gaussian distribution with zero mean and identity covariance; (iv) the data are generated with $p=10$ and $\theta_0\in\mathbb{R}^p$ such that $\theta_{0k}=1$ for all $k\in[p]$; and (v) a modified version of the seimparametric algorithm (i.e., Algorithm 2) was applied with $\gamma,\sigma$ chosen using cross-validation and $\Theta=[\frac{1}{2},2]$. The modification to Algorithm 2 is that we calculate $\tilde{x}_i=\min_x\{\|\bar{x}_i-x\|\mid x_k\geqslant 0\}$ and then compute $\hat{\theta}_n$ using SP–IOP–RISK–SAA, with the $\tilde{x}_i$ replacing the $\bar{x}_i$. The $\tilde{x}_i$ are the projection of the $\bar{x}_i$ onto the nonnegative orthant, and it turns out this projection does not affect our theoretical results. In particular, a short proof using the continuous mapping theorem (van der Vaart 2000) and the boundedness of the feasible set in R1 shows that $\max_{i\in[n]}\|\tilde{x}_i-S(u_i,\theta_0)\|\xrightarrow{p}0$. The projection is needed for this particular example because otherwise, the inverse formulation would contain logarithms of negative numbers, which are complex valued. More generally, a projection of $\bar{x}_i$ onto the feasible set of FOP will not affect our theoretical results and can be added as a step in our semiparametric algorithm.

In Table 3, we present estimation results for the first and second examples, averaged over 100 repetitions for each value of $n\in\{10,30,50,100,300,500,1{,}000\}$. We label the semiparametric algorithm (i.e., Algorithm 2) as SPA in the table. These results display estimation consistency of the semiparametric algorithm since it has lower estimation error as the data increase. To further illustrate estimation consistency, we conducted an experiment with the two situations above where the data were generated with $p=1$ and a $\theta_0$ that was

**Table 2.** Normalized Prediction Error $Q(\hat{\theta}_n)-\mathrm{var}(w)$ of Enumeration Algorithm (ENA) and Benchmark Algorithms (KKA and VIA) on Two Synthetic Instances ($n$ Increasing, $p=1$)

| | $n$ | 10 | 30 | 50 | 100 | 300 | 500 | 1,000 |
|---|---|---|---|---|---|---|---|---|
| Data: FOP-C | ENA | 0.0216 | 0.0184 | 0.0162 | 0.0150 | 0.0065 | 0.0046 | 0.0017 |
| Model: FOP-B | KKA | 0.0168 | 0.0124 | 0.0128 | 0.0151 | 0.0150 | 0.0150 | 0.0132 |
| | VIA | 0.0249 | 0.0185 | 0.0196 | 0.0149 | 0.0089 | 0.0072 | 0.0042 |
| Data: SQR-1 | ENA | 0.0294 | 0.0217 | 0.0152 | 0.0110 | 0.0073 | 0.0041 | 0.0024 |
| Model: FOP-B | KKA | 0.0394 | 0.0389 | 0.0398 | 0.0440 | 0.0504 | 0.0525 | 0.0518 |
| | VIA | 0.0343 | 0.0287 | 0.0243 | 0.0187 | 0.0122 | 0.0084 | 0.0072 |

**Table 3.** Estimation Error $\|\hat{\theta}_n - \theta_0\|$ of Semiparametric Algorithm (SPA) and Benchmark Algorithms (KKA and VIA) on Two Synthetic Instances ($n$ Increasing, $p = 10$)

| | $n$ | 10 | 30 | 50 | 100 | 300 | 500 | 1,000 |
|---|---|---|---|---|---|---|---|---|
| Data: FOP-D | SPA | 2.4618 | 1.7025 | 1.2543 | 0.8535 | 0.4754 | 0.3750 | 0.2573 |
| Model: FOP-D | KKA | 2.2569 | 1.5513 | 1.2229 | 0.9281 | 0.6107 | 0.5435 | 0.4447 |
| | VIA | 3.3829 | 3.2603 | 3.1937 | 3.1501 | 3.0292 | 3.0324 | 2.9208 |
| Data: FOP-E | SPA | 0.9189 | 0.7982 | 0.7500 | 0.7487 | 0.6639 | 0.6070 | 0.5783 |
| Model: FOP-E | KKA | 1.6687 | 1.5850 | 1.5813 | 1.5865 | 1.5828 | 1.5806 | 1.5811 |
| | VIA | 1.9299 | 1.6781 | 1.6826 | 1.6132 | 1.6001 | 1.5973 | 1.5843 |

randomly chosen from a uniform distribution with support $[0,1]$ and $[\frac{1}{2},2]$ for the first and second situations, respectively. A plot comparing the estimates $\hat{\theta}_n$ to the true parameter $\theta_0$ for the first situation when $n = 1{,}000$ is shown in Figure 2, and a plot comparing the estimates $\hat{\theta}_n$ to the true parameter $\theta_0$ for the second situation when $n = 1{,}000$ is shown in Figure 3. Consistent estimates should line up along the diagonal, and hence these plots demonstrate the estimation consistency (inconsistency) of the semiparametric algorithm (KKA and VIA). It is worth comparing the results of the semiparametric and enumeration algorithms. As mentioned above, the semiparametric algorithm will generally have a higher estimation error than the enumeration algorithm—this can be observed in these plots because the semiparametric algorithm estimates have a larger variation about the diagonal than the estimates of the enumeration algorithm.

In the second set of experiments, we generate data using a given model that is different than the FOP used to formulate SAA-IOP. In other words, this set of experiments includes situations where the model whose parameters are being identified does not match the model that generates the data. As a result, this setting consists of situations where IC is *not* satisfied. The first setting is where (i) the data are generated by FOP-C, which is $\min\{\frac{3}{2}\cdot x'x - (1+u)'x \,|\, x \in [0,1]^{10}\}$; (ii) the model estimated by IOP-SAA is FOP-D with $p = 10$; (iii) $u$ has a uniform distribution with support $[0,5]^{10}$; (iv) the measurement noise $w$ has a jointly Gaussian distribution with zero mean and identity covariance; and (v) the semiparametric algorithm (i.e., Algorithm 2) was applied with $\gamma, \sigma$ chosen using cross-validation (Hastie et al. 2009) and $\Theta = [0,2]$. The second setting is where (i) the data are generated by the statistical model SQR-P given by $y_i = \min\{\max\{\sqrt{u_i},0\},1\} + w_i$; (ii) the model estimated by IOP-SAA is FOP-D with $p = 10$; (iii) $u$ has a uniform distribution with support $[0,5]^{10}$; (iv) the measurement noise $w$ has a jointly Gaussian distribution with zero mean and identity covariance; and (v) the semiparametric algorithm (i.e., Algorithm 2) was applied with $\gamma$, $\sigma$ chosen using cross-validation (Hastie et al. 2009) and $\Theta = [0,2]$. The results averaged over 100 repetitions of sampling $n \in \{10, 30, 50, 100, 300, 500, 1{,}000\}$ data points and then estimating the parameter $\theta$ are summarized in Table 4, and these results are normalized by subtracting $\mathbb{E}(w'w)$. The reason for this normalization is that the prediction error $\mathbb{E}(\|y - \xi(u)\|^2)$ of the prediction $\xi(u)$ of the true model (either FOP-C or SQR-M, respectively) is $\mathbb{E}(w'w)$ because $y = \xi(u) + w$ here. The enumeration algorithm has a lower prediction error because it is risk consistent, whereas KKA and VIA are not risk consistent.

In the third set of experiments, we generate data using the previous four settings. The difference in this set of experiments is that we fix $n = 1{,}000$ and vary $p \in \{1, 3, 5, 10, 30\}$. The results when the data/model are given by FOP-D/FOP-D and FOP-E/FOP-E, averaged over 100 repetitions and then estimating the parameter $\theta$, are summarized in Table 5. These results show that the semiparametric algorithm has a lower

**Figure 3.** (Color online) Scatter Plot Comparing Estimated Parameter $\hat{\theta}_n$ vs. True Parameter $\theta_0$ as Computed by Different Algorithms at $n = 1{,}000$ When the Data and Model Are Both FOP-E

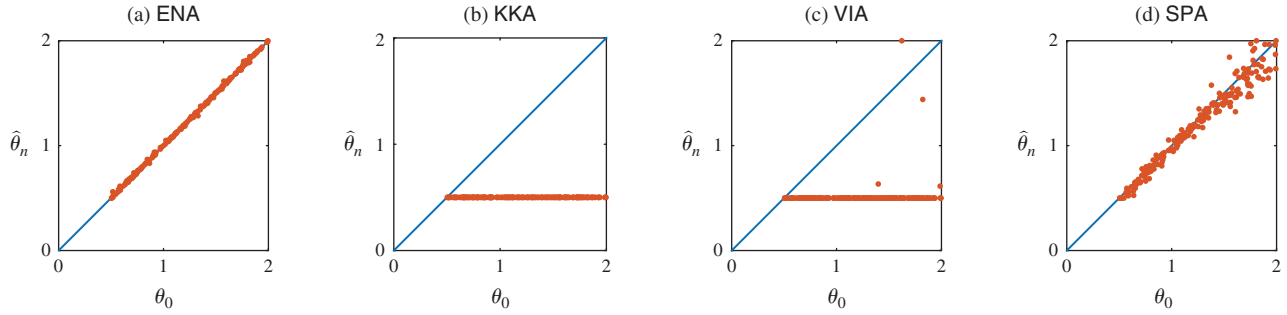


**Table 4.** Normalized Prediction Error $Q(\hat{\theta}_n) - \mathbb{E}(w'w)$ of Semiparametric Algorithm (SPA) and Benchmark Algorithms (KKA and VIA) on Two Synthetic Instances ($n$ Increasing, $p = 10$)

| | $n$ | 10 | 30 | 50 | 100 | 300 | 500 | 1,000 |
|---|---|---|---|---|---|---|---|---|
| Data: FOP-C | SPA | 0.2319 | 0.1972 | 0.1744 | 0.1501 | 0.1029 | 0.0844 | 0.0529 |
| Model: FOP-D | KKA | 0.1584 | 0.1308 | 0.1314 | 0.1349 | 0.1452 | 0.1497 | 0.1481 |
| | VIA | 0.3438 | 0.3407 | 0.3360 | 0.3205 | 0.2950 | 0.2816 | 0.2811 |
| Data: SQR-M | SPA | 0.4180 | 0.3497 | 0.3195 | 0.2470 | 0.1572 | 0.0998 | 0.0658 |
| Model: FOP-D | KKA | 0.3645 | 0.3885 | 0.3987 | 0.4537 | 0.5115 | 0.5114 | 0.5214 |
| | VIA | 0.3468 | 0.2784 | 0.2737 | 0.2524 | 0.2405 | 0.2458 | 0.2599 |

**Table 5.** Estimation Error $\|\hat{\theta}_n - \theta_0\|$ of Semiparametric Algorithm (SPA) and Benchmark Algorithms (KKA and VIA) on Two Synthetic Instances ($n = 1{,}000$, $p$ Increasing)

| | $p$ | 1 | 3 | 5 | 10 | 30 |
|---|---|---|---|---|---|---|
| Data: FOP-D | SPA | 0.0601 | 0.1464 | 0.1907 | 0.2794 | 0.4701 |
| Model: FOP-D | KKA | 0.1178 | 0.2349 | 0.3038 | 0.4619 | 0.7978 |
| | VIA | 0.4943 | 1.2254 | 1.8099 | 2.9522 | 5.7737 |
| Data: FOP-E | SPA | 0.0251 | 0.1258 | 0.2571 | 0.5890 | 0.5576 |
| Model: FOP-E | KKA | 0.5000 | 0.8660 | 1.1174 | 1.5804 | 2.7377 |
| | VIA | 0.5000 | 0.8691 | 1.1231 | 1.5966 | 2.7628 |

estimation error than KKA and VIA in these examples. The results when the data/model are given by FOP-C/FOP-B and SQR-M/FOP-B, averaged over 100 repetitions and then estimating the parameter $\theta$, are summarized in Table 6. These results show that the semiparametric algorithm hasa lower prediction error than KKA and VIA in these examples.

### 5.3. High-Dimensional Nonlinear Forward Problem with Stochastic Constraints

We now consider a setting where FOP is high dimensional, contains a logarithmic objective, and has an exponential stochastic constraint (i.e., the constraint depends on $u$). Specifically, we consider the following setting: (i) FOP-F is

$$\min\left\{-\sum_{k=1}^{p}\theta_k\cdot u_k^{(1)}\cdot\log(x_k)\,\middle|\,\frac{1}{p}\sum_{k=1}^{p}e^{x_k+u_k^{(1)}}-u_k^{(2)}\leqslant 0,\,x_k\geqslant 0\right\}, \tag{42}$$

**Table 6.** Normalized Prediction Error $Q(\hat{\theta}_n) - \mathbb{E}(w'w)$ of Semiparametric Algorithm (SPA) and Benchmark Algorithms (KKA and VIA) on Two Synthetic Instances ($n = 1{,}000$, $p$ Increasing)

| | $p$ | 1 | 3 | 5 | 10 | 30 |
|---|---|---|---|---|---|---|
| Data: FOP-C | SPA | 0.0064 | 0.0171 | 0.0403 | 0.0628 | 0.2048 |
| Model: FOP-D | KKA | 0.0538 | 0.1553 | 0.2619 | 0.5252 | 1.5712 |
| | VIA | 0.0078 | 0.0175 | 0.0745 | 0.2602 | 0.9654 |
| Data: SQR-M | SPA | 0.0056 | 0.0194 | 0.0319 | 0.0606 | 0.1568 |
| Model: FOP-D | KKA | 0.0148 | 0.0471 | 0.0761 | 0.1523 | 0.4394 |
| | VIA | 0.0055 | 0.0273 | 0.0821 | 0.2848 | 1.2896 |

(ii) $u^{(1)}$ has a uniform distribution with support $[1,2]^p$ and $u^{(2)}$ has a uniform distribution with support $[50,100]^p$, (iii) the measurement noise $w$ has a jointly Gaussian distribution with zero mean and identity covariance, (iv) the data are generated with $\theta_0 \in \mathbb{R}^p_+$ such that $\sum_{k=1}^{p}\theta_{0k} = p$, and (v) a modified version of the seimparametric algorithm (i.e., Algorithm 2) is applied where $\Theta = \{\theta \in \mathbb{R}^p_+ \mid \sum_{k=1}^{p}\theta_k = p\}$, and $\gamma, \sigma$ is selected using cross-validation. We set $n = 1{,}000$ and repeat the sampling and estimation procedure 100 times for each value of $p \in \{5, 10, 20, 50, 100\}$. The average estimation and prediction errors are summarized in Tables 7 and 8, respectively, which show that the semiparametric algorithm is competitive with existing methods in this setting as well. Note that the magnitude of the errors is expected to increase with $p$, since we do not normalize the error for the number of parameters being estimated.

### 5.4. Empirical Data: Estimating an Energy-Comfort Utility Function

We next apply our inverse optimization framework to the problem of estimating a utility function that

**Table 7.** Estimation Error $\|\hat{\theta}_n - \theta_0\|$ of Semiparametric Algorithm (SPA) and Benchmark Algorithms (KKA and VIA) on a Synthetic Instance ($n = 1{,}000$, $p$ Increasing)

| | $p$ | 5 | 10 | 20 | 50 | 100 |
|---|---|---|---|---|---|---|
| Data: FOP-F | SPA | 0.5535 | 0.8530 | 1.1522 | 2.0020 | 2.7205 |
| Model: FOP-F | KKA | 2.1753 | 4.6199 | 8.4599 | 12.4102 | 17.8112 |
| | VIA | 1.1825 | 1.8689 | 3.7320 | 5.9003 | 8.1874 |

**Table 8.** Normalized Prediction Error $Q(\hat{\theta}_n) - \mathbb{E}(w'w)$ of Semiparametric Algorithm (SPA) and Benchmark Algorithms (KKA and VIA) on a Synthetic Instance ($n = 1{,}000$, $p$ Increasing)

| | $p$ | 5 | 10 | 20 | 50 | 100 |
|---|---|---|---|---|---|---|
| Data: FOP-F | SPA | 0.2539 | 0.5117 | 0.9307 | 2.9423 | 5.8644 |
| Model: FOP-F | KKA | 8.2329 | 22.5149 | 60.4496 | 145.2210 | 302.1124 |
| | VIA | 2.3475 | 3.7495 | 10.8325 | 26.5713 | 48.3217 |

**Table 9.** Prediction Error $Q(\hat{\theta}_n)$ of Enumeration Algorithm (ENA) and Benchmark Algorithms (KKA and VIA) on a Temperature Preference Data Set ($n$ Increasing, $p$ Fixed)

| | $n$ | 10 | 30 | 50 | 100 | 300 | 500 | 1,000 |
|---|---|---|---|---|---|---|---|---|
| Data: SDH-E | ENA | 1.3656 | 1.3308 | 1.3255 | 1.3169 | 1.3112 | 1.3099 | 1.3090 |
| Model: FOP-S | KKA | 2.2439 | 2.2528 | 2.2508 | 2.2351 | 2.2225 | 2.2220 | 2.2200 |
| | VIA | 2.2975 | 2.2538 | 2.2472 | 2.2277 | 2.2163 | 2.2166 | 2.2138 |

describes the trade-off made between occupant comfort and the amount of energy consumption when setting a thermostat temperature set point for air conditioning. The data we use are collected from Sutardja Dai Hall on the Berkeley campus, which was used as part of the BRITE-S testbed in our past experiments (Aswani et al. 2012a, b, c) concerning robust learning-based optimization (Aswani et al. 2013) of heating, ventilation, and air conditioning systems. Specifically, this building is equipped with a commercial web application (Building Robotics 2016) that allows occupants to change the thermostat temperature set points in real time, and so the set points are changed throughout the year by occupants in response to factors such as the outside weather.

When a room is being cooled, a lower temperature set point requires increased energy consumption since the air conditioner must provide more cold air; however, the purpose of air conditioning is to improve comfort by lowering the room temperature. And so individuals must trade off comfort and energy consumption when choosing the set point. A simplified utility function model (expressed as minimization of the negative of the utility function) that captures this trade-off is FOP-S:

$$\min_x \{\theta_1 \cdot (x-76)^2 + (x-\theta_2-u)^2 \,|\, x \in [70,76]\}, \quad (43)$$

where $x \in \mathbb{R}$ is the thermostat temperature set point in units of degrees Fahrenheit (°F), and $u \in \mathbb{R}$ is the current outside temperaure in degrees Fahrenheit (°F). The term $(x-\theta_2-u)^2$ indicates a preference for a temperature set point that is a fixed amount $\theta_2$ above the outside temperature $u$ (i.e., the preferred temperature is $\theta_2+u$), and the reason for this term is that individuals prefer a higher indoor temperature as the outside temperature increases (American Society of Heating, Refrigeration, and Air-Conditioning Engineers 2013). The term $\theta_1 \cdot (x-76)^2$ indicates a preference for a higher set point because of energy considerations, and the number 76 is used because 76 °F–78 °F is a relatively high set point temperature that is often recommended for saving energy. The parameter $\theta_1$ quantifies the trade-off between the preference for a higher set point to save energy versus the desired indoor temperature $\theta_2+u$. Finally, the constraints $x \in [70,76]$ indicate observed set point limits.

The results averaged over 100 repetitions of sampling $n \in \{10,30,50,100,300,500,1{,}000\}$ data points and then estimating the parameters $\theta$ are summarized in Table 9. The data set (which we label SDH-E in the table) used consists of outside temperature measurements (i.e., the $u$ variable) and the chosen temperature set point (i.e., the $x$ variable) of a single thermostat in Sutardja Dai Hall. In each repetition, the full data set was randomly split into a 1,000-point *training* data set and a 14,500-point *testing* data set. The $n$ data points were randomly chosen from the training data set, and the prediction errors of the estimated parameters were computed using the testing data set. To evaluate the statistical significance of the computed results, a bootstrap hypothesis test (Efron and Tibshirani 1994) was conducted. The computed $p$-value was less than 0.01, which indicates that the improved performance of the enumeration algorithm is statistically significant.

## 6. Conclusion

We developed and analyzed a formulation for inverse optimization in the setting where noisy measurements of the optimal points of a convex optimization problem are available. Our approach requires solving a bilevel program, and we defined a new duality-based reformulation to convert this bilevel program into a single-level program. We showed that our formulation as a bilevel program leads to statistical consistency, in contrast to existing heuristics. Although our formulation is NP-hard to solve, we provided two numerical algorithms that maintain the statistical consistency of our formulation. Finally, we demonstrated that our approach improves on existing methods for inverse optimization through a series of numerical experiments using both synthetic and empirical data.

## Appendix A. Lemmas and Omitted Proofs

**Lemma 1.** *Suppose R4 holds. Then for $t > c_1 \cdot \gamma$, we have*

$$\mathbb{P}\left(\left|\gamma^{-m} \cdot \frac{1}{n}\sum_{j=1}^{n} K\left(\frac{u_j-u_i}{\gamma}\right) - \mu(u_i)\right| > t\right) \leqslant 2\exp(-2c_2 n\gamma^{2m} \cdot (t-c_1\cdot\gamma)^2), \quad (A.1)$$

*where $c_1, c_2 > 0$ are constants.*

**Proof.** Recall $\mu(u)$ is the probability density function of $u$, and note that

$$\begin{aligned}
&\left|\mu(u_i) - \mathbb{E}\left[\gamma^{-m} K\left(\frac{u-u_i}{\gamma}\right)\middle| u_i\right]\right| \\
&\quad = \left|\mu(u_i) - \gamma^{-m}\int_{\mathbb{R}^m} K\left(\frac{u-u_i}{\gamma}\right)\mu(u)\,du\right| \\
&\quad = \left|\mu(u_i) - \gamma^{-m}\int_{\mathbb{R}^m} K(s)\mu(u_i+\gamma s)\gamma^m\,ds\right| \\
&\quad = \left|\mu(u_i) - \int_{\mathbb{R}^m} K(s)(\mu(u_i)+\gamma\nabla\mu(u_i+\beta\gamma s)^T s)\,ds\right| \\
&\quad = \left|\int_{\mathbb{R}^m} K(s)\nabla\mu(u_i+\beta\gamma s)^T s\,ds\right|\cdot\gamma \leqslant c_1\cdot\gamma, \qquad \text{(A.2)}
\end{aligned}$$

where the second line follows from a change of variables $s = (u-u_i)/\gamma$, the third line follows from the multivariate form of Taylor's theorem with some $\beta\in[0,1]$, the fourth line follows because a kernel function has the property $\int K(u)\,du = 1$, and the fifth line follows by setting $c_1 = \max_{u\in U}|\int_{\mathbb{R}^m} K(s)\nabla\mu(u)^T s\,ds|$. Note this $c_1$ term is finite because (i) a kernel function has the property that its support is finite (i.e., $K(u)=0$ for $\|u\|>1$), and (ii) $\mu(u)$ is a continuously differentiable probability density function by R4. Next, note that by Hoeffding's inequality (Vershynin 2012), we have for $t>0$ that

$$\begin{aligned}
&\mathbb{P}\left(\left|\gamma^{-m}\cdot\frac{1}{n}\sum_{j=1}^n K\left(\frac{u_j-u_i}{\gamma}\right) - \mathbb{E}\left[\gamma^{-m}K\left(\frac{u-u_i}{\gamma}\right)\middle| u_i\right]\right| > t\right) \\
&\quad \leqslant 2\exp(-2c_2 n\gamma^{2m}t^2), \qquad \text{(A.3)}
\end{aligned}$$

where $c_2 = (\max_u K(u))^2$. Combining (A.2) and (A.3) gives the desired result. □

**Lemma 2.** *Suppose* A1 *and* R1–R4 *hold. Then for* $t > c_3\cdot\gamma^{1/2} + c_4\cdot\gamma$, *we have*

$$\begin{aligned}
&\mathbb{P}\left(\left\|\gamma^{-m}\cdot\frac{1}{n}\sum_{j=1}^n y_j\cdot K\left(\frac{u_j-u_i}{\gamma}\right) - \mu(u_i)S(u_i,\theta_0)\right\| > t\right) \\
&\quad \leqslant 2\exp(-2c_5 n\gamma^{2m}\cdot(t - c_3\cdot\gamma^{1/2} - c_4\cdot\gamma)), \qquad \text{(A.4)}
\end{aligned}$$

*where* $c_3, c_4, c_5 > 0$ *are constants.*

**Proof.** First, note that $S(u,\theta)$ consists of a single point from the strict convexity assumption in R3. Next, note that having A1 and R1–R4 means that proposition 4.41 of (Bonnans and Shapiro 2000) holds: This means for $\gamma>0$ sufficiently small, we have

$$\|S(u,\theta_0) - S(u_i,\theta_0)\| \leqslant \alpha\cdot\gamma^{1/2}, \qquad \text{(A.5)}$$

where $\alpha>0$ is a constant, whenever $\|u-u_i\|\leqslant\gamma$. Next, recall that $y_i$ conditioned on $u_i$ has distribution $S(u_i,\theta_0)+w_i$ under IC. Moreover, we have

$$\mathbb{E}\left[\gamma^{-m} y K\left(\frac{u-u_i}{\gamma}\right)\middle| u_i\right] = \mathbb{E}\left[\gamma^{-m} S(u,\theta_0) K\left(\frac{u-u_i}{\gamma}\right)\middle| u_i\right], \qquad \text{(A.6)}$$

since $\mathbb{E}(w_i)=0$ and $w_i$ is independent of $u_i$. Thus, we have

$$\begin{aligned}
&\left\|\mu(u_i)S(u_i,\theta_0) - \mathbb{E}\left[\gamma^{-m} y K\left(\frac{u-u_i}{\gamma}\right)\middle| u_i\right]\right\| \\
&\quad = \left\|\mu(u_i)S(u_i,\theta_0) - \gamma^{-m}\int_{\mathbb{R}^m} K\left(\frac{u-u_i}{\gamma}\right)\mu(u)S(u,\theta_0)\,du\right\| \\
&\quad = \left\|\mu(u_i)S(u_i,\theta_0) - \gamma^{-m}\int_{\mathbb{R}^m} K(s)\mu(u_i+\gamma s)\right. \\
&\qquad\qquad \left.\cdot S(u_i+\gamma s,\theta_0)\gamma^m\,ds\right\| \\
&\quad = \left\|\mu(u_i)S(u_i,\theta_0) - \int_{\mathbb{R}^m} K(s)(\mu(u_i)+\gamma\nabla\mu(u_i+\beta\gamma s)^T s)\right. \\
&\qquad\qquad \left.\cdot(S(u_i,\theta_0)+S(u_i+\gamma s,\theta_0)-S(u_i,\theta_0))\,ds\right\| \\
&\quad = \left\|\int_{\mathbb{R}^m} K(s)\mu(u_i)(S(u_i+\gamma s,\theta_0)-S(u_i,\theta_0))\,ds\right. \\
&\qquad \left. + \int_{\mathbb{R}^m} K(s)\gamma\nabla\mu(u_i+\beta\gamma s)^T s S(u,\theta_0)\,ds\right\| \leqslant c_3\cdot\gamma^{1/2} + c_4\cdot\gamma, \qquad \text{(A.7)}
\end{aligned}$$

where the second line follows from a change of variables $s = (u-u_i)/\gamma$, the third line follows from the multivariate form of Taylor's theorem with some $\beta\in[0,1]$, the fourth line follows because a kernel function has the property $\int K(u)\,du = 1$, and the fifth line follows from (A.5) and by setting $c_3 = \alpha\cdot\max_{u\in U}|\int_{\mathbb{R}^m} K(s)\mu(u)\,ds|$ and $c_4 = \max_{u\in U}(|\int_{\mathbb{R}^m} K(s)\nabla\mu(u)^T s\,ds|\cdot\|S(u,\theta_0)\|)$. Note the $c_3, c_4$ terms are finite because (i) a kernel function has the property that its support is finite (i.e., $K(u)=0$ for $\|u\|>1$), (ii) $\mu(u)$ is a continuously differentiable probability density function by R4, and (iii) $S(u,\theta_0)$ is bounded by R1. Next, note that $y$ is a subexponential random variable (Vershynin 2012) since (i) $S(u,\theta_0)$ is a bounded random variable by R1, and (ii) $w$ is subexponential by R4. Hence, by Hoeffding's inequality for subexponential random variables (Vershynin 2012), we have for $t>0$ that

$$\begin{aligned}
&\mathbb{P}\left(\left\|\gamma^{-m}\cdot\frac{1}{n}\sum_{j=1}^n y_j\cdot K\left(\frac{u_j-u_i}{\gamma}\right) - \mathbb{E}\left[\gamma^{-m} y K\left(\frac{u-u_i}{\gamma}\right)\middle| u_i\right]\right\| > t\right) \\
&\quad \leqslant 2\exp(-2c_5 n\gamma^{2m} t) \qquad \text{(A.8)}
\end{aligned}$$

for some $c_5>0$. Combining (A.7) and (A.8) gives the desired result. □

**Proof of Proposition 1.** We show this using a counterexample. Suppose FOP is $\min\{x^2 - (\theta+u)\cdot x \mid x\in[0,10]\}$, and note its solution set $S(u,\theta) = \min\{(u+\theta)/2, 10\}$ is single-valued. Assume the distribution of $u$ is

$$u = \begin{cases} 0 & \text{with probability (w.p.) } \frac{1}{2}, \\ 20 & \text{w.p. } \frac{1}{2}, \end{cases} \qquad \text{(A.9)}$$

and that the distribution of $w$ is

$$w = \begin{cases} -1 & \text{w.p. } \frac{1}{2}, \\ +1 & \text{w.p. } \frac{1}{2}. \end{cases} \qquad \text{(A.10)}$$

Finally, suppose $y = S(u,\theta)+w$, $\Theta = \{\theta\in\mathbb{R}\colon 0\leqslant\theta\leqslant 10\}$, and $\theta_0 = 10$. By construction, this problem satisfies A1, A2, and IC. Also, observe that the joint distribution of $(u,y)$ is

$$(u,y) = \begin{cases} (0,4), & \text{w.p. } \frac{1}{4}, \\ (0,6), & \text{w.p. } \frac{1}{4}, \\ (20,9), & \text{w.p. } \frac{1}{4}, \\ (20,11), & \text{w.p. } \frac{1}{4}. \end{cases} \qquad \text{(A.11)}$$

We show that both VIA and KKA are not estimation consistent for this problem.

We begin with VIA. This approach solves

$$\min_{\theta\in\Theta} \ \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$$
$$\text{s.t. } \nabla f(y_i,u_i,\theta)\cdot(x_i-y_i)\geqslant-\epsilon_i, \quad \forall\, x_i\in[0,10],\ \forall\, i\in[n]. \tag{A.12}$$

The constraint

$$\nabla f(y_i,u_i,\theta)\cdot(x_i-y_i)\geqslant-\epsilon_i, \quad \forall\, x_i\in[0,10] \tag{A.13}$$

is a variational inequality, and VIA exactly reformulates this using linear duality. We operate with the original variational inequality since the reformulation in VIA is exact and does not change the solution. If $y_i=4$, then a straightforward calculation shows that (A.13) is equivalent to one of the constraints: $\epsilon_i\geqslant 4\cdot(8-\theta)$ if $\theta\leqslant 8$ or $\epsilon_i\geqslant -6\cdot(8-\theta)$ if $\theta>8$. If $y_i=6$, then (A.13) is equivalent to the constraint $\epsilon_i\geqslant 6\cdot(12-\theta)$. If $y_i=9$, then (A.13) is equivalent to the constraint $\epsilon_i\geqslant 2+k$. Finally, if $y_i=11$, then (A.13) is equivalent to one of the following constraints: $\epsilon_i\geqslant 11\cdot(2-\theta)$ if $\theta\leqslant 2$ or $\epsilon_i\geqslant 2-\theta$ if $\theta>2$. Next, we solve the problem $\min\{\epsilon_i^2\,|\,(A.13)\}$ for each possible value of $y_i$ and $\theta$. If $y_i=4$, then the minimum is $16\cdot(8-\theta)^2$ if $\theta\leqslant 8$ and $36\cdot(8-\theta)^2$ if $\theta>8$. If $y_i=6$, then the minimum is $36\cdot(12-\theta)^2$. If $y_i=9$, then the minimum is $(2+\theta)^2$. If $y_i=11$, then the minimum is $121\cdot(2-\theta)^2$ if $\theta\leqslant 2$ and 0 if $\theta>2$. Thus, we have

$$4\cdot\mathbb{E}(\epsilon_i^2)=\begin{cases}36\cdot(12-\theta)^2+(2+\theta)^2+121\cdot(2-\theta)^2\\ \quad+16\cdot(8-\theta)^2 & \text{if } \theta\leqslant 2,\\ 36\cdot(12-\theta)^2+(2+\theta)^2+16\cdot(8-\theta)^2 & \\ \quad\text{if } \theta\in(2,8], & \\ 36\cdot(12-\theta)^2+(2+\theta)^2+36\cdot(8-\theta)^2 & \\ \quad\text{if } \theta>8. & \end{cases} \tag{A.14}$$

Finally, we solve the optimization problem $\min\{\mathbb{E}(\epsilon_i^2)\,|\,\theta\in[0,10]\}$. A simple calculation shows that the minimum occurs at $\theta^*=718/73\approx 9.8356$. However, the minimizer of (A.12) will converge in probability to $\theta^*$, because (i) we can exactly reformulate (A.12) as

$$\min_{\theta\in\Theta} \ \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$$
$$\text{s.t. } \epsilon_i^2=\begin{cases}16\cdot(8-\theta)^2\cdot\mathbb{1}(\theta\leqslant 8)+36\cdot(8-\theta)^2 & \\ \quad\cdot\mathbb{1}(\theta>8) & \text{if } y_i=4,\\ 36\cdot(12-\theta)^2 & \text{if } y_i=6,\\ (2+\theta)^2 & \text{if } y_i=9,\\ 121\cdot(2-\theta)^2\cdot\mathbb{1}(\theta\leqslant 2) & \text{if } y_i=11,\end{cases} \quad \forall\, i\in[n]; \tag{A.15}$$

which implies that (ii) we can apply the uniform law of large numbers (Jennrich 1969), since $\epsilon_i^2$ as defined in (A.15) is a continuous function; thus (iii) we get convergence of the minimizer from a standard consistency result in statistics (see, for instance, van der Vaart 2000, theorem 5.7 or Bickel and Doksum 2006, theorem 5.2.3). This shows VIA is not estimation consistent, since $\theta_0=10$.

Next, we consider KKA. This approach solves

$$\min_{\theta\in\Theta} \ \frac{1}{n}\sum_{i=1}^{n}\|\epsilon_i\|^2$$
$$\begin{aligned}\text{s.t. } &\nabla f(y_i,u_i,\theta)-\lambda_{i1}+\lambda_{i2}=\epsilon_{i1}\\ &-\lambda_{i1}\cdot y_i=\epsilon_{i2},\\ &\lambda_{i2}\cdot(y_i-10)=\epsilon_{i3},\\ &\lambda_i\geqslant 0.\end{aligned} \tag{A.16}$$

We first solve the problem (A.16), with $n=1$, for each possible value of $y_i$ and $\theta$. If $y_i=4$, then the minimum is $(16/17)\cdot(8-\theta)^2$ if $\theta\leqslant 8$ and $(36/37)\cdot(8-\theta)^2$ if $\theta>8$. If $y_i=6$, then the minimum is $(36/37)\cdot(12-\theta)^2$. If $y_i=9$, then the minimum is $\frac{1}{2}\cdot(2+\theta)^2$. If $y_i=11$, then the minimum is $(121/122)\cdot(2-\theta)^2$ if $\theta\leqslant 2$ and $\frac{1}{2}\cdot(2-\theta)^2$ if $\theta>2$. Thus, we have

$$4\cdot\mathbb{E}(\|\epsilon_i\|^2)=\begin{cases}\frac{36}{37}\cdot(12-\theta)^2+\frac{1}{2}\cdot(2+\theta)^2+\frac{121}{122}\cdot(2-\theta)^2\\ \quad+\frac{16}{17}\cdot(8-\theta)^2 & \text{if } \theta\leqslant 2,\\ \frac{36}{37}\cdot(12-\theta)^2+\frac{1}{2}\cdot(2+\theta)^2+\frac{1}{1}\cdot(2-\theta)^2 & \\ \quad+\frac{16}{17}\cdot(8-\theta)^2 & \text{if } \theta\in(2,8],\\ \frac{36}{37}\cdot(12-\theta)^2+\frac{1}{2}\cdot(2+\theta)^2+\frac{1}{2}\cdot(2-\theta)^2 & \\ \quad+\frac{36}{37}\cdot(8-\theta)^2 & \text{if } \theta>8.\end{cases} \tag{A.17}$$

Finally, we solve the optimization problem $\min\{\mathbb{E}(\|\epsilon_i\|^2)\,|\,\theta\in[0,10]\}$. A simple calculation shows that the minimum occurs at $\theta^*=12{,}080/1{,}833\approx 6.5903$. However, the minimizer of (A.16) will converge in probability to $\theta^*$, because we can exactly reformulate (A.16) as

$$\min_{\theta\in\Theta} \ \frac{1}{n}\sum_{i=1}^{n}\|\epsilon_i\|^2$$
$$\text{s.t. } \|\epsilon_i\|^2=\begin{cases}\frac{16}{17}\cdot(8-\theta)^2\cdot\mathbb{1}(\theta\leqslant 8)+\frac{36}{37} & \\ \quad\cdot(8-\theta)^2\cdot\mathbb{1}(\theta>8), & \text{if } y_i=4,\\ \frac{36}{37}\cdot(12-\theta)^2 & \text{if } y_i=6,\\ \frac{1}{2}\cdot(2+\theta)^2 & \text{if } y_i=9,\\ \frac{121}{122}\cdot(2-\theta)^2\cdot\mathbb{1}(\theta\leqslant 2) & \\ \quad+\frac{1}{2}\cdot(2-\theta)^2 & \text{if } y_i=11,\end{cases} \quad \forall\, i\in[n], \tag{A.18}$$

which implies that we can apply the uniform law of large numbers (Jennrich 1969) since $\|\epsilon_i\|^2$ as defined in (A.18) is a continuous function; thus we get convergence of the minimizer from a standard consistency result in statistics (see, for instance, van der Vaart 2000, theorem 5.7 or Bickel and Doksum 2006, theorem 5.2.3). This shows that KKA is not estimation consistent, since $\theta_0=10$. □

**Proof of Corollary 2.** It suffices to show that risk consistency is necessary for estimation consistency in the counterexample given in the proof of Proposition 1. First note that the risk function

$$\begin{aligned}Q(\theta)&=\mathbb{E}\left(\left\|y-\min\left\{\frac{u+\theta}{2},10\right\}\right\|^2\right)\\ &=\frac{1}{4}\cdot\left(\left(4-\frac{\theta}{2}\right)^2+\left(6-\frac{\theta}{2}\right)^2+(9-10)^2+(11-10)^2\right)\end{aligned} \tag{A.19}$$

is continuous since $\Theta = \{\theta \in \mathbb{R}: 0 \leqslant \theta \leqslant 10\}$. Now suppose a sequence $\hat{\theta}_n$ is estimation consistent. Since $\hat{\theta}_n \xrightarrow{p} \theta_0$, by continuity of $Q(\theta)$ and the continuous mapping theorem (van der Vaart 2000), we have $Q(\hat{\theta}_n) \xrightarrow{p} Q(\theta_0)$. Since $\arg\min\{Q(\theta) \,|\, \theta \in \Theta\} = 10 = \theta_0$, and $\hat{\theta}_n \to \theta_0$, we have that $\hat{\theta}_n$ converges to a minimizer of $Q(\theta)$. Hence $\hat{\theta}_n$ is risk consistent. □

## Appendix B. Identifiability in Inverse Optimization

Estimation consistency in any statistical setting (including inverse optimization with noisy data) requires that an identifiability condition holds, and such identifiability conditions can be stated under a variety of different mathematical formulations (Wald 1949, Jennrich 1969, Bartlett and Mendelson 2002, Greenshtein and Ritov 2004, Bickel and Doksum 2006, Chatterjee 2014, Aswani 2016). The intuition for these different formulations is the same: in essence, an identifiability condition states that the output of the model is different for two distinct sets of model parameters. It is important to note that identifiability is a statistical property of the model and the error metric used. Consequently, it is possible for an estimator to be statistically inconsistent, even when an identifiability condition holds (see, for instance, Proposition 1). In the context of inverse optimization with noisy data, we define an identifiability condition IC.

Showing that IC holds is complicated by the presence of constraints in FOP. To illustrate this, consider two related instances of FOP with $x \in \mathbb{R}$ and $\theta \in [0,2]$. The first, $\min(x-\theta)^2$, is FOP-I, and the second, $\min\{(x-\theta)^2 \,|\, x \leqslant 1\}$, is FOP-II. Since these two problems are strictly convex, their minimizers are unique. Next, suppose we would like to estimate $\theta$ given a (noiseless) measurement $y_i$ of the minimizer. Observe that FOP-I is identifiable because we must have $\theta = y_i$. However, FOP-II is not identifiable because if $y_i = 1$, then we may have any $\theta \in [1,2]$. Thus, the constraint $x \leqslant 1$ renders FOP-II unidentifiable and precludes the possibility of IC holding for FOP-II.

Although FOP-II is not identifiable, a related problem is identifiable because of external inputs. In particular, consider the forward problem $\min\{(x-\theta-u)^2 \,|\, x \leqslant 1\}$, where $x \in R$ and $\theta \in [0,2]$. This problem is strictly convex, and so its minimizer is unique for each fixed value of $u$. In fact, the minimizer is given by $y_i = \min\{(\theta + u_i), 1\}$. And so a sufficient condition for identifiability of FOP-III is if $\mathbb{P}(u_i \leqslant -1) > 0$. For instance, if $u_i = -1$, then $y_i = \theta - 1$, and so $\theta$ is uniquely determined by $y_i$. The presence of the input parameter $u$ ensures identifiability of FOP-III.

## Endnotes

[1] R1 can be relaxed to requiring a nonempty *relative* interior if the affine constraints of FOP are of the form $Mx + \zeta(u,\theta) = 0$, where $M$ is a matrix and $\zeta$ is a continuous function. The reason is that our proofs make use of a result (Rockafellar and Wets 1998, example 5.10) on the continuity of parametrized convex constraints with a nonempty interior, and this result can be generalized for the above case through minor modifications (using corresponding results on relative interiors from Rockafellar and Wets 1998, section 2.H) to ensure continuity of the feasible set of FOP with a nonempty relative interior. Generalizing Rockafellar and Wets (1998, example 5.10) or our results to cases with more complex affine constraints will require further study.

[2] Note that this notion of near-optimality is defined with respect to IOP–SAA, whereas the definition of near-optimality given in (15) is with respect to the regularized formulation R–IOP–SAA.

[3] Technically, this theorem applies to maximizing upper semicontinuous functions, but the results and proof trivially extend to the case of minimizing lower semicontinuous functions.

## References

Aalami HA, Parsa Moghaddam M, Yousefi GR (2010) Demand response modeling considering interruptible/curtailable loads and capacity market programs. *Applied Energy* 87(1):243–250.

Ahuja RK, Orlin JB (2001) Inverse optimization. *Oper. Res.* 49(5): 771–783.

American Society of Heating, Refrigeration, and Air-Conditioning Engineers (2013) Thermal environmental conditions for human occupancy. ANSI/ASHRAE Standard 55-2013, American Society of Heating, Refrigeration, and Air-Conditioning Engineers, Atlanta.

Aswani A (2016) Low-rank approximation and completion of positive tensors. *SIAM J. Matrix Anal. Appl.* 37(3):1337–1364.

Aswani A, Tomlin C (2012) Incentive design for efficient building quality of service. *Allerton Conf. Comm., Control, Comput.* (IEEE, Piscataway, NJ), 90–97.

Aswani A, Gonzalez H, Sastry S, Tomlin C (2013) Provably safe and robust learning-based model predictive control. *Automatica* 49(5):1216–1226.

Aswani A, Kaminsky P, Mintz Y, Flowers E, Fukuoka Y (2016) Predictive modeling of behavior in weight loss interventions. Working paper, University of California, Berkeley, Berkeley.

Aswani A, Master N, Taneja J, Krioukov A, Culler D, Tomlin C (2012a) Energy-efficient building HVAC control using hybrid system LBMPC. *IFAC Proc. Vol.* 45(17):496–501.

Aswani A, Master N, Taneja J, Krioukov A, Culler D, Tomlin C (2012b) Quantitative methods for comparing different HVAC control schemes. *Internat. Conf. Performance Evaluation Methodologies Tools* (IEEE, Piscataway, NJ), 326–332.

Aswani A, Master N, Taneja J, Smith V, Krioukov A, Culler D, Tomlin C (2012c) Identifying models of HVAC systems using semiparametric regression. *Proc. Amer. Control Conf.* (IEEE, Piscataway, NJ), 3675–3680.

Audet C, Hansen P, Jaumard B, Savard G (1997) Links between linear bilevel and mixed 0–1 programming problems. *J. Optim. Theory Appl.* 93(2):273–300.

Bajari P, Benkard CL, Levin J (2007) Estimating dynamic models of imperfect competition. *Econometrica* 75(5):1331–1370.

Bard JF, Moore JT (1990) A branch and bound algorithm for the bilevel programming problem. *SIAM J. Scientific Statist. Comput.* 11(2):281–292.

Bartlett P, Mendelson S (2002) Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Machine Learn. Res.* 3(November):463–482.

Beil DR, Wein LM (2003) An inverse-optimization-based auction mechanism to support a multiattribute rfq process. *Management Sci.* 49(11):1529–1545.

Berge C (1963) *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity* (Courier Dover Publications, Mineola, NY).

Bertsimas D, Gupta V, Paschalidis IC (2012) Inverse optimization: A new perspective on the Black-Litterman model. *Oper. Res.* 60(6):1389–1403.

Bertsimas D, Gupta V, Paschalidis IC (2015) Data-driven estimation in equilibrium using inverse optimization. *Math. Programming* 153(2):595–633.

Bickel P, Doksum K (2006) *Mathematical Statistics: Basic Ideas and Selected Topics*, 2nd ed., Vol. 1 (Pearson/Prentice Hall, Upper Saddle River, NJ).

Bonnans JF, Ioffe A (1995a) Quadratic growth and stability in convex programming problems with multiple solutions. *J. Convex Anal.* 2(1–2):41–57.

Bonnans JF, Ioffe A (1995b) Second-order sufficiency and quadratic growth for nonisolated minima. *Math. Oper. Res.* 20(4):801–817.

Bonnans J, Shapiro A (2000) *Perturbation Analysis of Optimization Problems* (Springer, New York).

Boyd S, Vandenberghe L (2009) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Building Robotics (2016) Comfy home page. Accessed June 1, 2016, http://www.comfyapp.com.

Burton D, Toint PhL (1992) On an instance of the inverse shortest paths problem. *Math. Programming* 53(1–3):45–61.

Carr S, Lovejoy W (2000) The inverse newsvendor problem: Choosing an optimal demand portfolio for capacitated resources. *Management Sci.* 46(7):912–927.

Chan TCY, Craig T, Lee T, Sharpe MB (2014) Generalized inverse multiobjective optimization with application to cancer therapy. *Oper. Res.* 62(3):680–695.

Chatterjee S (2014) A new perspective on least squares under convex constraint. *Ann. Statist.* 42(6):2340–2381.

Crama P, Reyck BDe, Degraeve Z (2008) Milestone payments or royalties? Contract design for R&D licensing. *Oper. Res.* 56(6): 1539–1552.

Dempe S, Kalashnikov V, Pérez-Valdés G, Kalashnikova N (2015) *Bilevel Programming Problems* (Springer, Berlin).

Efron B, Tibshirani RJ (1994) *An Introduction to the Bootstrap*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability (Taylor & Francis, Abingdon, UK).

Erkin Z, Bailey MD, Maillart LM, Schaefer AJ, Roberts MS (2010) Eliciting patients' revealed preferences: An inverse Markov decision process approach. *Decision Anal.* 7(4):358–365.

Faragó A, Szentesi Á, Szviatovszki B (2003) Inverse optimization in high-speed networks. *Discrete Appl. Math.* 129(1):83–98.

Green PE, Srinivasan V (1990) Conjoint analysis in marketing: New developments with implications for research and practice. *J. Marketing* 54(4):3–19.

Greenshtein E, Ritov Y (2004) Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* 10(6):971–988.

Hastie T, Tibshirani R, Friedman J (2009) *Elements of Statistical Learning*, 2nd ed. (Springer, New York).

Haviv I, Regev O (2012) Tensor-based hardness of the shortest vector problem to within almost polynomial factors. *Theory Comput.* 8(1):513–531.

Heuberger C (2004) Inverse combinatorial optimization: A survey on problems, methods, and results. *J. Combinatorial Optim.* 8(3): 329–361.

Hillar C, Lim L-H (2013) Most tensor problems are NP-hard. *J. ACM* 60(6):45:1–45:39.

Hochbaum DS (2003) Efficient algorithms for the inverse spanning-tree problem. *Oper. Res.* 51(5):785–797.

Iyengar G, Kang W (2005) Inverse conic programming with applications. *Oper. Res. Lett.* 33(3):319–330.

Jennrich RI (1969) Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* 40(2):633–643.

José Fortuny-Amat BM (1981) A representation and economic interpretation of a two-level programming problem. *J. Oper. Res. Soc.* 32(9):783–792.

Keshavarz A, Wang Y, Boyd S (2011) Imputing a convex objective function. 2011 *IEEE Internat. Sympos. Intelligent Control* (*ISIC*) (IEEE, Piscataway, NJ), 613–619.

Ratliff LJ, Dong R, Ohlsson H, Sastry SS (2014) Incentive design and utility learning via energy disaggregation. *IFAC Proc. Vol.* 47(3): 3158–3163.

Rockafellar RT, Wets RJ-B (1998) *Variational Analysis*, Vol. 317 (Springer, Berlin).

Saez-Gallego J, Morales JM, Zugno M, Madsen H (2016) A data-driven bidding model for a cluster of price-responsive consumers of electricity. *IEEE Trans. Power Systems* 31(6):5001–5011.

Schaefer AJ (2009) Inverse integer programming. *Optim. Lett.* 3(4): 483–489.

Tao T (2012) *Topics in Random Matrix Theory*, Graduate Studies in Mathematics, Vol. 132 (American Mathematical Society, Providence, RI).

Troutt MD, Pang W-K, Hou S-H (2006) Behavioral estimation of mathematical programming objective function coefficients. *Management Sci.* 52(3):422–434.

Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. *Science* 211(4481):453–458.

van der Vaart AW (2000) *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, UK).

Vershynin R (2012) Introduction to the non-asymptotic analysis of random matrices. Eldar YC, Kutyniok G, eds. *Compressed Sensing: Theory and Applications* (Cambridge University Press, Cambridge, UK), 210–268.

Wald A (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* 20(4):595–601.

Wang L (2009) Cutting plane algorithms for the inverse mixed integer linear programming problem. *Oper. Res. Lett.* 37(2):114–116.

Zhang H, Zenios S (2008) A dynamic principal-agent model with hidden information: Sequential optimality through truthful state revelation. *Oper. Res.* 56(3):681–696.

Zhang J, Liu Z (1996) Calculating some inverse linear programming problems. *J. Computational Appl. Math.* 72(2):261–273.

Zhang J, Xu C (2010) Inverse optimization for linearly constrained convex separable programming problems. *Eur. J. Oper. Res.* 200(3):671–679.

**Anil Aswani** is an assistant professor in the Department of Industrial Engineering and Operations Research at the University of California, Berkeley. His research interests include data-driven decision making, with particular emphasis on addressing inefficiencies and inequities in health systems and physical infrastructure.

**Zuo-Jun (Max) Shen** is a Chancellor's Professor in the Department of Industrial Engineering and Operations Research and the Department of Civil and Environmental Engineering at University of California, Berkeley. He is also an honorary professor at Tsinghua University. He has been active in the following research areas: integrated supply chain design and management, design and analysis of optimization algorithms, energy system and transportation system planning, and optimization.

**Auyon Siddiq** is an assistant professor in the Anderson School of Management at the University of California, Los Angeles. His research interests include data analytics, operations management, and policy issues.