INFORMATION BASED INFERENCE IN MODELS WITH SET-VALUED PREDICTIONS AND MISSPECIFICATION

HIROAKI KAIDO
Department of Economics, Boston University

FRANCESCA MOLINARI
Department of Economics, Cornell University

This paper proposes an information-based inference method for partially identified parameters in incomplete models that is valid both when the model is correctly specified and when it is misspecified. Key features of the method are: (i) it is based on minimizing a suitably defined Kullback-Leibler information criterion that accounts for incompleteness of the model and delivers a non-empty pseudo-true set; (ii) it is computationally tractable; (iii) its implementation is the same for both correctly and incorrectly specified models; (iv) it exploits all information provided by variation in discrete and continuous covariates; (v) it relies on Rao's score statistic, which is shown to be asymptotically pivotal.

KEYWORDS: Misspecification, Partial Identification, Rao's score statistic.

1. INTRODUCTION

Applied research is rarely based on the empirical evidence alone: exogeneity assumptions, behavioral restrictions, distributional and functional form specifications, etc., are routinely imposed to approximate features of complex social and economic phenomena.

Hiroaki Kaido: hkaido@bu.edu

Francesca Molinari: fm72@cornell.edu

We thank the Editor and the anonymous referees, and Xiaoxia Shi and seminar participants at Bonn/Mannheim, Bristol/Warwick, BU, CEMFI, Chicago, Columbia, Cornell, JHU, Michigan, Nebraska, NYU, Queen's Mary, São Paulo School of Economics, Tokyo, Toulouse, UCD, UCL, UCLA, UCSB, UPF, USC, Yale, Wisconsin, ESAM21, AMES23, ESWC25, Workshop on Econometrics and Models of Strategic Interaction, and Chamberlain Seminar, for comments. Undral Byambadalai, Shuowen Chen, Qifan Han, Luis Hoderlein, Yan Liu, Yiqi Liu, Patrick Power, Ryan Strong, Yiwei Sun provided excellent research assistance. We gratefully acknowledge financial support from NSF grants SES-2018498 (Kaido) and 1824375 (Molinari).

Early on, Koopmans and Reiersol (1950, p. 169) highlighted the importance of imposing restrictions based on prior knowledge of the phenomenon under analysis and some criteria of simplicity, but argued against choosing these restrictions primarily for the purpose of identifiability of a parameter that the researcher happens to be interested in. Yet, even after embracing Koopmans and Reiersol's perspective, concerns for misspecification remain.

In response to these concerns, we propose a novel information theoretic inference method in the spirit of White (1982) that is asymptotically valid both when the model is correctly or incorrectly specified and point or partially identified, with the latter case often resulting when researchers remain agnostic about features of the model where they lack prior knowledge. Our method is easy to implement and addresses three challenges that arise in misspecified partially identified models: (i) the set of observationally equivalent parameters may be spuriously tight or empty; (ii) confidence sets constructed assuming correct model specification may (severely) undercover; (iii) their tightness may be misinterpreted as highly informative data.[1] The method applies to a specific but wide class of partially identified models that predict a set of values for the endogeneous variables (Y) given the exogenous observed and unobserved ones (X and U, respectively), yielding a set of conditional distributions for $Y|X$. Many examples belong to this class, including: games with multiple equilibria; discrete choice models with either interval data on covariates, counterfactual choice sets, endogeneous explanatory variables, or unobserved heterogeneity in choice sets; dynamic discrete choice models; network formation models; and auctions and school choice models under weak assumptions on behavior (see, e.g., Molinari, 2020).

We adapt the textbook method for (point identified) models that predict a singleton conditional distribution for $Y|X$, to models that predict a set of distributions for $Y|X$, through three steps that jointly yield our main innovations. In step one, leveraging a result in Artstein (1983), we characterize the exact set of model predicted distributions consistent with all maintained assumptions,[2] and show that with discrete Y it is a finite dimensional convex

[1]Counterpart challenges arise in point identified models, with various solutions put forward since at least White (1982) (see, e.g., Hansen and Lee, 2021, and references therein for a recent discussion).

[2]Using only a subset of model implications may yield misleading conclusions (Li et al., 2024, Molinari, 2020, Beresteanu et al., 2011); nonetheless, even in this case our method remains applicable.

polytope. Doing so greatly simplifies our next steps as it allows us to bypass the need to model all possible distribution of $Y|X$ through the use of probability mixtures with infinite dimensional nuisance mixing functions (an approach pursued, e.g., in Chen et al., 2011, 2018, under the assumption of correct specification). In step two, we define a never-empty pseudo-true set, denoted Θ^*, for the parameter vector θ characterizing the model. This is the collection of minimizers of a Kullback-Leibler (KL) information criterion measuring the divergence of the set of model-predicted distributions from the distribution of the observed data. The set Θ^* shrinks to the pseudo-true parameter vector in White (1982) if the modeling assumptions are augmented so that the model predicts a single distribution. When the model predicts multiple distributions, Θ^* collects the parameter values that minimize the researcher's ignorance about the true structure (Akaike, 1973, White, 1982), recognizing that this ignorance extends to the selection mechanism that picks an element from the set of model predictions. As in the point identified case, one may wonder to what extent a pseudo-true set is of substantive interest. In our view, models are only approximations to the true data generating process (DGP) and hence pseudo-true sets are often what researchers estimate in practice; hence, inference methods robust to misspecification are needed.

To this end, in the third step we obtain a profiled likelihood function by projecting, with respect to the KL divergence measure, the distribution of the observed data on the set of model implied distributions. A key advantage, yielded by steps one and two, is that this projection is carried out through a computationally simple convex program, which in our leading examples with discrete outcomes features a strictly convex objective and linear constraints. As in the textbook case, the pseudo-true set equals the collection of maximizers of the profiled likelihood function. We next derive a novel score representation for this function, based on d_θ estimating (score) equations, with d_θ the number of model's parameters. These equations depend on the conditional distribution of $Y|X$, which is unknown and needs to be estimated nonparametrically. We leverage classic results in the semiparametric inference literature, specifically Newey (1994), to establish that an orthogonality property holds. Provided the convergence rate of the nonparametric conditional density estimator is sufficiently fast ($o_p(n^{-1/4})$), this implies that the limit distribution of the averaged score function is insensitive to estimation of the distribution of the data. We use this result to construct a Rao's score statistic with asymptotically pivotal limit distribution $\chi^2_{d_\theta}$, which we use to test the hypothesis that a candidate parameter vector belongs to Θ^*. We invert the test to construct a confidence set and show that it is robust to misspecification: it covers each element of Θ^* with asymptotic probability at least equal to the nominal level $1-\alpha$, uniformly over a large class of DGPs. While the size of Θ^* may or may not be impacted by the extent to which the model is misspecified (see Section 3.4), relative to Θ^*'s size the volume of the confidence set depends only on the sampling variability of the score statistic.

Related Literature. Chen et al. (2011, 2018) put forward inference methods that asymptotically cover the $\arg\max$ of a profiled likelihood function, which by textbook arguments is related to the $\arg\min$ of the KL divergence that we focus on. However, their method relies on infinite dimensional nuisance functions to represent each possible distribution of $Y|X$, and its validity is established for correctly specified models only. Our method completely bypasses the use of infinite dimensional nuisance functions, yielding computational advantages and allowing us to consider broader classes of incomplete models.[3]

Only a few recent papers put forth tools for construction of confidence sets that are valid in the presence of misspecification, covering each element of a pseudo-true identified set with an asymptotic probability at least as large as a prespecified nominal level. Andrews and Kwon (2024) show that model misspecification can lead to spuriously tight confidence sets while statistical tests have low power at detecting misspecification. They propose a notion of pseudo true set, a specification test, and an asymptotically uniformly valid inference method for partially identified models defined by a finite number of unconditional moment inequalities. Their test statistic aggregates violations of the sample moment conditions relaxed by the minimum amount that guarantees that at least one parameter vector in the parameter space satisfies them. Stoye (2020) studies interval identified scalar parameters with asymptotic normality of the estimators of the endpoints. He obtains a valid and never-empty confidence interval that is free of tuning parameters and simple to compute. In

[3]See the discussion following (3.9) for further comparisons. For a dynamic model where misspecification results when subjective beliefs deviate from rational expectations, Chen et al. (2024) show that a confidence set built using the results in Chen et al. (2018) is asymptotically valid.

contrast, our method applies to models for conditional density functions of outcome variables given discrete and continuous covariates. Allowing for the latter is important: they are commonplace in practice and may yield substantial identifying information. While a fine discretization or the use of instrument functions (e.g., Andrews and Shi, 2013) may transform conditional moment inequalities in unconditional ones, doing so may incur computational costs associated with an increase in the number of moment inequalities proportional to the cardinality of the discretized support or the number of instrument functions. We bypass this problem by evaluating directly the contribution of each observation to the score function. Our pseudo-true set and inference method are insensitive to which inequalities one uses to characterize the sharp collection of model implied distributions for $Y|X$. In comparison, much of the related literature often requires moment selection either for computational tractability or as part of the inference procedure, which may substantially impact the population region that the researcher targets (Li et al., 2024) and the properties of confidence sets (e.g., Andrews and Shi, 2013, Bugni et al., 2017, Kaido et al., 2019).

Outline. Section 2 introduces the class of models we study. Section 3 provides the notion of pseudo-true set and derives the misspecification robust inference method. Section 4 discusses computational aspects of the method. Section 5 provides an empirical illustration revisiting the analysis in Kline and Tamer (2016) and Section 6 Monte Carlo evidence on the size and power of the test. Section 7 concludes. Appendix A provides proofs of our main results. The Online Appendix includes auxiliary Lemmas and additional examples.

2. NOTATION AND MOTIVATING EXAMPLE

Let $Y \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$, $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ and $U \in \mathcal{U} \subseteq \mathbb{R}^{d_U}$ denote, respectively, observable endogenous and exogenous variables, and unobservable variables, with realizations y, x, u. Let $P_0 \in \mathcal{P}(\mathcal{Y} \times \mathcal{X})$ denote the distribution of (Y, X).[4] Assume the conditional law $P_0(\cdot|x)$ is absolutely continuous with respect to a σ-finite measure μ on $\mathcal{Y}$. Let $p_{0,y|x}$ be the Radon-Nikodym derivative of $P_0(\cdot|x)$ with respect to μ and $p_0 \equiv \{p_{0,y|x}, x \in \mathcal{X}\}$. To simplify exposition, let p_0 be in the interior of the unit simplex. We consider a framework where,

[4] For a space $\mathcal{S}$ with Borel σ-algebra $\Sigma_{\mathcal{S}}$, $\mathcal{P}(\mathcal{S})$ denotes the set of all Borel probability measures on $(\mathcal{S}, \Sigma_{\mathcal{S}})$.

based on prior knowledge, a researcher is willing to maintain some parametric restrictions on the joint behavior of (Y,X,U) and posit that these continue to hold, or change according to the model's specification, under a counterfactual intervention. The researcher, however, is agnostic about other features of the model or whether they continue to hold after the intervention. To accommodate this framework, denoting $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ the parameter vector characterizing the model, we let the structure associate with each (u,x,θ) a set of predicted outcomes through a closed-valued and measurable correspondence $G:\mathcal{U}\times\mathcal{X}\times\Theta\mapsto\mathcal{Y}$.[5] This framework nests as a special case the textbook model with singleton predictions, where $Y=g(U|X;\theta)$ a.s. for $g:\mathcal{U}\times\mathcal{X}\times\Theta\mapsto\mathcal{Y}$ a measurable function. We assume the family of distributions for the latent variables U is known up to finite dimensional parameter vector that is part of θ, and, omitting specific notation for subvectors of θ, we denote $\{F_\theta:\theta\in\Theta\}$ the family of distributions for U and assume that F_θ is independent of X.[6] While assuming a parametric distribution for the entire vector U simplifies our presentation, in some applications the researcher might be unwilling to take a stand on the distribution of some of its components. Our method is applicable in these cases too, though it requires a parametric distribution for a subvector of U, as we show in Online Appendix Example C.2 for the case of panel dynamic discrete choice models, where the researcher may lack prior knowledge of the distribution of the initial condition (e.g., Honoré and Tamer, 2006).

The next example, used throughout the paper to illustrate results, clarifies notation. More examples are provided in the Online Appendix and in Molinari (2020).

Example 1 (Static entry game)**.** Consider a two player entry game as in Tamer (2003), with each player $i=1,2$ choosing to enter ($Y_i=1$) or stay out of the market ($Y_i=0$). Let (X_1,X_2) and $(U_1,U_2)\sim F_\theta$ be, respectively, observable and unobservable payoff shifters and player's payoffs be $\pi_j=Y_j(X_j\beta_j+\delta_jY_{(3-j)}+U_j), j=1,2,$ with $\delta_1\le 0,\delta_2\le 0$ the interaction effects and $(\beta_1,\beta_2,\delta_1,\delta_2)$ part of θ. Let each player enter the market if and only if $\pi_j\ge 0$. Given $\theta\in\Theta$ and $x\in\mathcal{X}$, the model has multiple pure strategy Nash equilibria (PSNE), depicted in Figure 1 as a function of (u_1,u_2). In our notation, the set of PSNE is

[5]Given a probability space $(\Omega,\mathfrak{F},\mathbf{P})$ and $\mathcal{C}$ the family of closed sets in $\mathbb{R}^d$, a correspondence $G:\Omega\mapsto\mathcal{C}$ is measurable if, for every compact set K in $\mathbb{R}^d$, $G^{-1}(K)=\{\omega\in\Omega:G(\omega)\cap K\neq\emptyset\}\in\mathfrak{F}$.

[6]This can easily be relaxed if the researcher is willing to specify the conditional distribution of $U|X$.

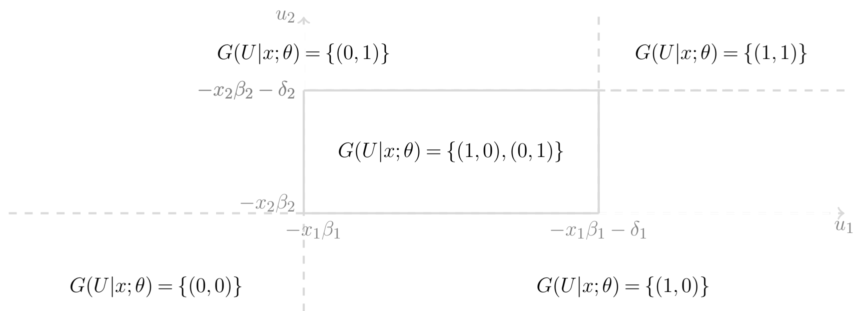

FIGURE 1.—Stylized depiction of $G(\cdot|x;\theta)$ in Example 1 with $\delta_1<0,\delta_2<0$.

the measurable correspondence $G(\cdot|x;\theta)$ (Beresteanu et al., 2011, Proposition 3.1), with:

$$G(U|x;\theta)=\{(0,0)\} \text{ if } U\in S_{\{(0,0)\}|x;\theta}\equiv\{u:u_j<-x_j\beta_j,\ j=1,2\}, \tag{2.1}$$

$$G(U|x;\theta)=\{(1,1)\} \text{ if } U\in S_{\{(1,1)\}|x;\theta}\equiv\{u:u_j\geq-x_j\beta_j-\delta_j,\ j=1,2\}, \tag{2.2}$$

$$\begin{aligned}G(U|x;\theta)=\{(1,0)\} \text{ if } U\in S_{\{(1,0)\}|x;\theta}&\equiv\{u:u_1\geq-x_1\beta_1-\delta_1,u_2<-x_2\beta_2-\delta_2)\}\\&\bigcup\{u:-x_1\beta_1\leq u_1<-x_1\beta_1-\delta_1,u_2<-x_2\beta_2\},\end{aligned} \tag{2.3}$$

$$\begin{aligned}G(U|x;\theta)=\{(0,1)\} \text{ if } U\in S_{\{(0,1)\}|x;\theta}&\equiv\{u:u_1<-x_1\beta_1,u_2\geq-x_2\beta_2\}\\&\bigcup\{u:-x_1\beta_1\leq u_1<-x_1\beta_1-\delta_1,u_2\geq-x_2\beta_2-\delta_2\},\end{aligned} \tag{2.4}$$

$$G(U|x;\theta)=\{(1,0),(0,1)\} \text{ if } U\in M_{x;\theta}\equiv\{u:-x_j\beta_j\leq u_j<-x_j\beta_j-\delta_j,\ j=1,2\}. \tag{2.5}$$

If one assumes $\delta_1\times\delta_2=0$ (a "principal assumption" in the econometrics literature on simultaneous equation models with dummy endogeneous variables), the region $M_{x;\theta}$ occurs with probability zero, and $G(U|x;\theta)$ reduces to a measurable function $g(U|x;\theta)$. Doing so, however, removes the simultaneity in player's actions that many applications aim to capture. One may consider completing the model through a selection mechanism that picks an outcome in the region of multiplicity, $M_{x;\theta}$. Doing so is also often undesirable, as various plausible selection mechanisms may lead to notably different conclusions about the nature of firms' competition (e.g. Berry, 1992, Table VII), and which firms enters when the market can sustain only one profitable entrant ($U\in M_{x;\theta}$) may be impacted by counterfactual interventions. The specification of F_θ can be made flexible through the use of mixtures, although our approach requires the mixture to be finite. □

3. INFORMATION-BASED INFERENCE ROBUST TO MISSPECIFICATION

3.1. *The Set of Model-Implied Density Functions*

One can view $G(U|x;\theta)$ as the collection of its *measurable selections* (Molchanov and Molinari, 2018, Def. 2.1), i.e., all random vectors $\tilde{Y}$ such that $\tilde{Y} \in G(U|x;\theta)$ a.s. Each selection $\tilde{Y}$ is a model predicted outcome. In order to obtain a set-valued analog of a likelihood model, one needs to be able to characterize the distribution of each of these predicted outcomes. To do so in a computationally feasible manner, we denote by $\mathcal{C}$ the collection of closed subsets of $\mathcal{Y}$ and define the law of $G(U|x;\theta)$ induced by the model's structure:

$$\nu_\theta(A|x) \equiv \int_{\mathcal{U}} \mathbf{1}(G(u|x;\theta) \subseteq A) dF_\theta(u), \quad \forall A \in \mathcal{C}. \tag{3.1}$$

The *containment functional* $\nu_\theta(A|x)$ uniquely determines the distribution of $G(U|x;\theta)$ when it is evaluated at all $A \in \mathcal{C}$ (Molchanov and Molinari, 2018, p.20).

Given $\theta \in \Theta$, $x \in \mathcal{X}$, and $\nu_\theta(\cdot|x)$, by Artstein (1983, Theorem 2.1) it is possible to characterize all distributions of measurable selections of $G(U|x;\theta)$ as the set

$$\text{core}(\nu_\theta(\cdot|x)) \equiv \{Q \in \mathcal{M}(\Sigma_Y, \mathcal{X}) : Q(A|x) \geq \nu_\theta(A|x), A \subseteq \mathcal{C}\}, \tag{3.2}$$

where $\mathcal{M}(\Sigma_Y, \mathcal{X})$ is the collection of laws of random variables supported on $\mathcal{Y}$ conditional on X. The characterization in (3.2) is sharp, in the sense that, up to an ordered coupling (Molchanov and Molinari, 2018, Chapter 2), given $\tilde{Y} \sim Q(\cdot|x)$, $\tilde{Y} \in G(U|x;\theta)$ a.s. if and only if $Q(A|x) \geq \nu_\theta(A|x)$, for all $A \subseteq \mathcal{C}$, x-a.s.

Example 1 (Continued). Let $Y_{x;\theta}(u)$ be the unique element of $G(u|x;\theta)$ if $u \notin M_{x;\theta}$ and (arbitrary) $Y_{x;\theta}(u) \equiv (0,0)$ if $u \in M_{x;\theta}$. Molchanov and Molinari (2018, Example 2.6) show that all measurable selections of $G(\cdot|x;\theta)$ in Example 1 can be represented as

$$Y(U,R) = Y_{x;\theta}(U)\mathbf{1}(U \notin M_{x;\theta}) + (R \times (0,1) + (1-R) \times (1,0))\mathbf{1}(U \in M_{x;\theta}), \tag{3.3}$$

for a random variable $R \in \{0,1\}$ with any distribution in $\mathcal{P}(\{0,1\})$ and unrestricted dependence on U given X. Each of these distributions is a *selection mechanism* (e.g., Ciliberto and Tamer, 2009) that assigns to $(1,0)$ and $(0,1)$ the probability that each is played given $X=x$ and $U \in M_{x;\theta}$. Beresteanu et al. (2011, Lemma 2.1) show that the distribution of each selection in (3.3) belongs to $\mathrm{core}(\nu_\theta(\cdot|x))$, and that only those distributions do. The containment functional of $G(\cdot|x;\theta)$ satisfies: $\nu_\theta(\{(0,0)\}|x) = F_\theta(S_{\{(0,0)\}|x;\theta})$, $\nu_\theta(\{(0,1),(1,0)\}|x) = 1 - F_\theta(S_{\{(0,0)\}|x;\theta}) - F_\theta(S_{\{(1,1)\}|x;\theta})$, and similarly for all $A \subseteq \mathcal{Y} = \{(0,0),(1,0),(0,1),(1,1)\}$, where for a given set $B \subset \mathcal{U}$, $F_\theta(B) = \int_{\mathcal{U}} \mathbf{1}(u \in B) dF_\theta(u)$ and the sets $S_{\{y\}|x;\theta}, y \in \mathcal{Y}$ are defined in (2.1)-(2.4). $\square$

Assume that there are σ-finite measures μ on $(\mathcal{Y}, \Sigma_Y)$ and ξ on $(\mathcal{X}, \Sigma_X)$, a product measure $\zeta \equiv \mu \times \xi$ on $(\mathcal{Y} \times \mathcal{X}, \Sigma_Y \times \Sigma_X)$, and for all $\theta \in \Theta$, $x \in \mathcal{X}$, and $Q \in \mathrm{core}(\nu_\theta(\cdot|x))$, $Q \ll \mu$ (this requirement is typically unrestrictive; see, e.g., White, 1982, p. 2). Let the set of conditional densities associated with $\mathrm{core}(\nu_\theta(\cdot|x))$ be

$$\mathfrak{q}_{\theta,x} \equiv \{q_{y|x} : q_{y|x} = dQ(\cdot|x)/d\mu,\ Q \in \mathrm{core}(\nu_\theta(\cdot|x))\}, \tag{3.4}$$

$$\mathfrak{q}_\theta \equiv \left\{\mathfrak{q}_{\theta,x},\ x \in \mathcal{X}\right\}. \tag{3.5}$$

Example 1 (Continued). Given $\theta \in \Theta$ and denoting Δ the unit simplex in $\mathbb{R}^4$, the set of all model predicted probability mass functions corresponding to selections of $G(\cdot|x;\theta)$ is

$$\begin{aligned} \mathfrak{q}_\theta = \Big\{ q_{y|x} \in \Delta :\ & q_{y|x}((0,0)|x) = F_\theta(S_{\{(0,0)\}|x;\theta}); \qquad q_{y|x}((1,1)|x) = F_\theta(S_{\{(1,1)\}|x;\theta}); \\ & F_\theta(S_{\{(1,0)\}|x;\theta}) \le q_{y|x}((1,0)|x) \le F_\theta(S_{\{(1,0)\}|x;\theta}) + F_\theta(M_{x;\theta}), \quad x \in \mathcal{X} \Big\}, \end{aligned} \tag{3.6}$$

with $S_{\{(0,0)\}|x;\theta}$, $S_{\{(1,1)\}|x;\theta}$, $S_{\{(1,0)\}|x;\theta}$, $M_{x;\theta}$ defined in (2.1), (2.2), (2.3), (2.5). $\square$

3.2. *Correct Specification, Misspecification, and Pseudo-True Set*

Define a model $\mathfrak{Q} \equiv \{\mathfrak{q}_\theta : \theta \in \Theta\}$ as the collection of sets $\mathfrak{q}_\theta$ across $\theta \in \Theta$. We propose a generalization of the standard definition of correct specification for models with singleton predictions (e.g., White, 1996, Def. 2.5) to models with set-valued predictions.

DEFINITION 3.1—Correctly Specified Model & Misspecified Model: *A model is correctly specified if $p_0 \in \mathfrak{q}_\theta$ for some $\mathfrak{q}_\theta \in \mathfrak{Q} \equiv \{\mathfrak{q}_\vartheta : \vartheta \in \Theta\}$, and misspecified otherwise.*

REMARK 3.1: *In models that yield a singleton prediction $Y = g(U|X;\theta)$ a.s., with $g : \mathcal{U} \times \mathcal{X} \times \Theta \mapsto \mathcal{Y}$ a measurable function, there is a unique implied law for $g|X = x$: $Q_\theta(A|x) = \int_{\mathcal{U}} \mathbf{1}(g(u|x;\theta) \in A) dF_\theta(u),\ \forall A \in \mathcal{C}$, with associated conditional density function $q_{\theta,y|x} = dQ_\theta(\cdot|x)/d\mu$ (compare with (3.1) and (3.4)). The model is defined as the collection of (singleton) $q_{\theta,y|x}$ across $\theta \in \Theta$ and $x \in \mathcal{X}$, $\mathcal{Q} = \left\{[q_{\theta,y|x},\ x \in \mathcal{X}] :\ \theta \in \Theta\right\}$. The model is correctly specified if $p_0 = q_\theta$ for some $q_\theta \in \mathcal{Q}$, and misspecified otherwise.*

Given density functions f and f', it has been common to measure their similarity through the *Kullback-Leibler Information Criterion (KLIC)* (White, 1996). Here we extend the standard definition of KLIC to measure divergence from f of a set of density functions $\mathfrak{f}$.

DEFINITION 3.2—KLIC for set of density functions: *Let $(\Omega, \mathfrak{F}, \zeta)$ be a measure space. Let $f : \Omega \mapsto \mathbb{R}_+$ be a measurable function satisfying $\int f d\zeta < \infty$ and $\int_S f \ln f d\zeta < \infty$ where $S = \{\omega \in \Omega : f(\omega) > 0\}$. Let $\mathfrak{f}$ denote a* set *of measurable functions $f' : \Omega \mapsto \mathbb{R}_+$ satisfying $\int_S f \ln f' d\zeta < \infty$ and $I(f||f') \equiv \int_S f \ln \frac{f}{f'} d\zeta$. The Kullback-Leibler divergence measure from f of a set $\mathfrak{f}$ is $I(f||\mathfrak{f}) \equiv \inf_{f' \in \mathfrak{f}} I(f||f')$.*

It follows from White (1996, Theorem 2.3) that when $\inf_{f' \in \mathfrak{f}} \int_S (f - f') d\zeta \geq 0$, $I(f||\mathfrak{f}) = 0$ if $f \in \mathfrak{f}$, and $I(f||\mathfrak{f}) > 0$ otherwise. As our approach is based on measuring divergence between conditional density functions, denoted $f(y|x)$ and $f'(y|x)$, in $I(f||\mathfrak{f})$ we replace the unconditional KLIC with the conditional one, $I(f||f') \equiv \int_{\mathcal{Y} \times \mathcal{X}} f(y,x) \ln \frac{f(y|x)}{f'(y|x)} d\zeta(y,x)$. Following White (1982)'s treatment of point-identified models, we let the *pseudo true set*, $\Theta^*(p_0)$, be the set of minimizers of the researcher's ignorance about the true structure. But here one is also ignorant about which selection from the model predicted set is closest to the data. Hence, minimization occurs with respect to both $\vartheta \in \Theta$, as in the textbook case, and $q \in \mathfrak{q}_\vartheta$. If the model is correctly specified, $\Theta^*(p_0)$ equals the sharp identification region, as in point-identified models where the pseudo-true value matches the data-generating one.

DEFINITION 3.3: *The* pseudo-true identified set *is given by*

$$\Theta^*(p_0) \equiv \left\{ \theta \in \Theta : I(p_0||\mathfrak{q}_\theta) = \inf_{\vartheta \in \Theta} I(p_0||\mathfrak{q}_\vartheta) \right\}. \tag{3.7}$$

To understand the effect of minimizing KLIC with respect to $q \in \mathfrak{q}_\vartheta$, note that

$$I(p_0||\mathfrak{q}_\vartheta) = \inf_{q \in \mathfrak{q}_\vartheta} \int_{\mathcal{Y}\times\mathcal{X}} p_0(y,x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\zeta(y,x)$$
$$= \int_{\mathcal{X}} p_{0,x}(x) \inf_{q_{y|x} \in \mathfrak{q}_{\vartheta,x}} \int_{\mathcal{Y}} p_{0,y|x}(y|x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\mu(y) d\xi(x), \tag{3.8}$$

with $\mathfrak{q}_{\vartheta,x}$ defined in (3.5). No unknown selection mechanisms are used in (3.8) to formalize all ways in which a measurable selection could be picked from $G(\cdot|x;\theta)$ and all associated likelihoods obtained. Rather, (3.8) relies on a convex program, with strictly convex objective and convex constraints (a finite dimensional convex program with linear constraints if $\mathcal{Y}$ is finite). It delivers the density function in $\mathfrak{q}_{\vartheta,x}$ closest with respect to KLIC to $p_{0,y|x}$,

$$q^*_{\vartheta,y|x} = \arg \inf_{q_{y|x} \in \mathfrak{q}_{\vartheta,x}} \int_{\mathcal{Y}} p_{0,y|x}(y|x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\mu(y), \tag{3.9}$$

which can be calculated analytically or numerically.[7] It can be interpreted as a profiled (quasi)-likelihood where a convex optimization program profiles out the selection mechanism, which is left completely unspecified and may arbitrarily depend on (X, U, ϑ). The support of X is also unrestricted. In contrast, related likelihood-based inference methods rely on an infinite-dimensional parameter space to represent the selection mechanism that picks measurable selections from $G(\cdot|x;\theta)$ (as in (3.3)) and profile it out via non-convex optimization programs with increasing number of (sieve) coefficients (e.g., Chen et al., 2011); or restrict the class of selection mechanisms by assuming that they do not depend on U after conditioning on X and that X has finite support (Chen et al., 2018).[8] Doing so may substantially increase computational burden or narrow the class of models allowed for. For example, in discrete choice models with unobserved heterogeneity in choice sets (Barseghyan et al., 2021) it would rule out choice set formation based on sequential search or rational inattention (see Online Appendix Example C.1).

[7] Existence and uniqueness of $q^*_{\vartheta,y|x}$ is guaranteed under mild conditions (see Online Appendix Lemma B.2).

[8] Chen et al. (2011)'s inference method assumes correct model specification. Chen et al. (2018, Remark 3) suggest that their method may remain valid for some misspecified separable models with discrete covariates.

Putting together (3.8) and (3.9) we obtain

$$I(p_0||\mathfrak{q}_\vartheta) = \int_{\mathcal{Y}\times\mathcal{X}} p_0(y,x) \ln \frac{p_{0,y|x}(y|x)}{q^*_{\vartheta,y|x}(y|x)} d\zeta(y,x).$$

Hence, the pseudo-true set $\Theta^*(p_0)$ in (3.7) is equal to the set of maximizers of $L(\vartheta)$, with

$$L(\vartheta) \equiv \mathbb{E}_{p_0}\left[L(\theta|X)\right] \text{ and } L(\theta|X) \equiv \mathbb{E}[\ln q^*_{\theta,y|x}(Y|X)|X]. \tag{3.10}$$

Example 1 (Continued). Given $\theta \in \Theta$, $x \in \mathcal{X}$, and $S_{\{(0,0)\}|x;\theta}$, $S_{\{(1,1)\}|x;\theta}$, $S_{\{(1,0)\}|x;\theta}$, $M_{x;\theta}$ as in (2.1), (2.2), (2.3), (2.5), let

$$\eta_1(\theta;x) \equiv 1 - F_\theta(S_{\{(0,0)\}|x;\theta}) - F_\theta(S_{\{(1,1)\}|x;\theta}), \tag{3.11}$$

$$\eta_2(\theta;x) \equiv F_\theta(S_{\{(1,0)\}|x;\theta}) + F_\theta(M_{x;\theta}), \tag{3.12}$$

$$\eta_3(\theta;x) \equiv F_\theta(S_{\{(1,0)\}|x;\theta}). \tag{3.13}$$

In words, $\eta_1(\theta;x)$ is the probability allocated by the model to either $(1,0)$ or $(0,1)$ occurring as outcome of the game; $\eta_2(\theta;x)$ [$\eta_3(\theta;x)$] is the upper [lower] bound implied by the model on the probability that $(1,0)$ is the outcome of the game. Define the parameter sets:

$$\Theta_1(x,p_0) \equiv \left\{\theta \in \Theta : \eta_3(\theta;x) \le \tfrac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) \le \eta_2(\theta;x)\right\} \tag{3.14}$$

$$\Theta_2(x,p_0) \equiv \left\{\theta \in \Theta : \tfrac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) > \eta_2(\theta;x)\right\} \tag{3.15}$$

$$\Theta_3(x,p_0) \equiv \left\{\theta \in \Theta : \tfrac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) < \eta_3(\theta;x)\right\}. \tag{3.16}$$

Then the profiled likelihood is given by (see Proposition B.1 in the Online Appendix):

$$q^*_{\theta,y|x}((0,0)|x) = F_\theta(S_{\{(0,0)\}|x;\theta}) \tag{3.17}$$

$$q^*_{\theta,y|x}((1,1)|x) = F_\theta(S_{\{(1,1)\}|x;\theta}) \tag{3.18}$$

$$q^*_{\theta,y|x}((0,1)|x) = \begin{cases} \frac{p_{0,y|x}((0,1)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) & \theta \in \Theta_1(x,p_0) \\ \eta_1(\theta;x) - \eta_2(\theta;x) & \theta \in \Theta_2(x,p_0) \\ \eta_1(\theta;x) - \eta_3(\theta;x) & \theta \in \Theta_3(x,p_0) \end{cases} \tag{3.19}$$

$$q^*_{\theta,y|x}((1,0)|x) = 1 - q^*_{\theta,y|x}((0,0)|x) - q^*_{\theta,y|x}((1,1)|x) - q^*_{\theta,y|x}((0,1)|x). \tag{3.20}$$

Intuitively, when $\theta \in \Theta_1(x, p_0)$, $(1,0)$ can be assigned a share of $\eta_1(\theta; x)$ equal to the share of $p_{0,y|x}(\{(0,1),(1,0)\}|x)$ that $(1,0)$ has in the data. When $\theta \in \Theta_2(x, p_0)$, that allocation yields a probability for $(1,0)$ larger than the model's upper bound $\eta_2(\theta; x)$, and the KL divergence is minimized setting $q^*_{\theta,y|x}((1,0)|x) = \eta_2(\theta; x)$. Similarly for $\theta \in \Theta_3(x, p_0)$. In Section 3.4 below we discuss the general topological properties of Θ^*. □

3.3. *The Score Function*

Here we characterize the *score function* associated with the singleton-valued likelihood function in (3.10), under the following regularity conditions.

ASSUMPTION 1: *(a) $\mathcal{Y}$ is finite and $\mathcal{X}$ is compact. (b) There is a collection $\mathcal{A}_G(x) \subset 2^{\mathcal{Y}}$, that does not depend on θ, such that* $\text{supp}(G(\cdot|x;\theta)) \equiv \{A \subseteq \mathcal{Y} : F_\theta(G(U|x;\theta) = A) > 0\} = \mathcal{A}_G(x)$ *for all $\theta \in \Theta$, $P_0 - a.s.$ (c) $\nu_\theta(A|X)$ is continuously differentiable with respect to θ and $\nabla_\theta \nu_\theta(A|X)$ is square integrable for all $A \subset \mathcal{Y}$, $P_0 - a.s.$ (d) $\Theta^*(p_0) \subset \text{int}\,\Theta$. (e) There exists a constant $c > 0$ such that for all $\theta \in \Theta$ and $y \in \mathcal{Y}$, $q^*_{\theta,y|x}(y|x) > c$, $P_0 - a.s.$*

Assumption 1-(a) restricts attention to models with discrete outcomes and compact support for X. Part (b) requires the support of $G(\cdot|x, \theta)$ not to vary with $\theta \in \Theta$, though it can vary with $x \in \mathcal{X}$. It is used in Lemma B.2 to show that the Lagrange multiplier vector for the convex program in (3.9) is unique and hence $L(\theta|X)$ is differentiable. Conditions that imply this uniqueness can replace it. Assumption 1-(b) holds in several applications, including discrete choice models with unobserved heterogeneity in choice sets (Barseghyan et al., 2021, see our Online Appendix C) and dynamic monopoly entry models (Berry and Compiani, 2022, as shown in Luo et al., 2025, p.17). Moreover, Gu et al. (2025, Remark 3.1) show that with finite $\mathcal{Y}$, Θ can be finitely partitioned in x-dependent partitions, so that within each partition $\mathcal{A}_G(x)$ does not depend on θ and all our results apply within the partition. Nonetheless, the assumption is restrictive as one crosses over from one partition to the other. For example, as stated it rules out that Θ includes both values of θ at which

$G(\cdot|X;\theta)$ collapses to a function and values at which it is a non-singleton correspondence (whether the latter values of θ are consistent with the DGP may be testable, see Chen and Kaido, 2023). In Online Appendix D we show that we can dispense with this condition and work, along crossing surfaces between the partitions, with the subdifferential of $L(\theta|X)$, but doing so makes our procedure less tractable as further explained following Theorem 3.4. Assumption 1-(c) is easily verified when F_θ is differentiable in θ. Part (d) implies that first order conditions hold at all $\theta^* \in \Theta^*(p_0)$. Part (e) bounds $q^*_{\theta,y|x}$ away from zero. We verify all conditions for Example 1 below and other examples in Online Appendix C.

THEOREM 3.1: *(i) Under Assumption 1(a)(b)(c), $L(\theta|X)$ is differentiable with respect to θ on $\mathrm{int}(\Theta)$, $P_0 - a.s.$ (ii) Under Assumption 1, there exists a function $s : \Theta \times \mathcal{Y} \times \mathcal{X} \times \Delta \to \mathbb{R}^{d_\theta}$, with Δ the unit-simplex in $\mathbb{R}^{|\mathcal{Y}|-1}$, such that $\mathbb{E}[\|s_\theta(Y|X;p_{0,y|x})\|^2] < \infty$, and*

$$\frac{\partial}{\partial\theta} L(\theta|x) = \mathbb{E}[s_\theta(Y|X;p_{0,y|x})|X = x], \tag{3.21}$$

$$\mathbb{E}[s_\theta(Y|X;p_{0,y|x})] = 0, \;\; \text{for all } \theta \in \Theta^*(p_0). \tag{3.22}$$

The score function depends on $p_{0,y|x}$. When $\theta \mapsto L(\theta|x)$ is concave, $\Theta^*(p_0)$ equals the set of $\theta \in \Theta$ for which (3.22) holds. When concavity does not hold, this set includes $\Theta^*(p_0)$. As one of our goals is to avoid spuriously tight confidence sets, we view the benefit of an easy-to-implement method to outweigh the cost of a sometimes wider confidence set which asymptotically uniformly covers the set of θ's satisfying (3.22). Our Monte Carlo results in Section 6 show that, for the examples analyzed there, our procedure performs well relative to existing methods. The proof of Theorem 3.1 is in Appendix A and leverages results in Gauvin and Janin (1990) to establish differentiability with respect to θ of

$$L(\theta|x) = \mathbb{E}[\ln q^*_{\theta,y|x}(Y|X)|X = x] = \sup_{q_{y|x} \in \mathfrak{q}_{\theta,x}} \sum_{y \in \mathcal{Y}} p_{0,y|x}(y|x) \ln q_{y|x}(y|x), \tag{3.23}$$

where $\mathfrak{q}_{\theta,x}$ is defined in (3.4). The proof uses results in Luo et al. (2025) by which under Assumption 1-(b), the smallest collection of inequalities among the ones in (3.4) that suffice to sharply characterize $\mathfrak{q}_{\theta,x}$ does not depend on θ. As further discussed in Section 3.6, Lemma B.2-(ii) in the Online Appendix shows that $q^*_{\theta,y|x}$ and the value of $L(\theta|x)$, and hence its differentiability and the score function $s_\theta(y|x;p_{0,y|x})$, are insensitive to inclusion of additional inequalities from (3.4) in the maximization problem in (3.23). Hence, so are the pseudo-true set $\Theta^*(p)$ and our inference procedure. This is in contrast with much related literature, where moment selection is often required for computational tractability or for the inference procedure, but may have substantial implications both on the population region that the researcher targets (Li et al., 2024) and on the properties of the inference procedure (e.g., Andrews and Shi, 2013, Bugni et al., 2017, Kaido et al., 2019).

Example 1 (Continued). Assumption 1-(a) holds. Part (b) holds if, e.g., $\delta_1, \delta_2 < 0$, in which case $\mathcal{A}_G = \{\{(0,0)\}, \{(0,1)\}, \{(1,0)\}, \{(1,1)\}, \{(1,0),(0,1)\}\}$ for all $\theta \in \Theta$. If one allows for $\delta_1 = \delta_2 = 0$, at that value the model is complete and the support of $G(\cdot|x,\theta)$ changes to $\{\{(0,0)\}, \{(0,1)\}, \{(1,0)\}, \{(1,1)\}\}$, violating Assumption 1-(b); in that case, one can extend our approach using subdifferentials, as we show in Online Appendix D. Assumption 1-(c) holds as long as $F_\theta(S_{y|x;\theta})$, $y \in \mathcal{Y}$, and $F_\theta(M_{x;\theta})$ are differentiable with respect to θ (e.g., for (U_1, U_2) bivariate normal). Part (e) follows because one can find $c > 0$ such that $\eta_j(\theta;x) \geq c$, $j = 1, \ldots, 3$ and $\eta_1(\theta;x) - \eta_j(\theta;x) \geq c$, $j = 2, 3$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$. In Proposition B.1 in the Online Appendix we show that:

$$s_\theta((0,0)|x;p_{0,y|x}) = \frac{\nabla_\theta F_\theta(S_{\{(0,0)\}|x;\theta})}{F_\theta(S_{\{(0,0)\}|x;\theta})} \tag{3.24}$$

$$s_\theta((1,1)|x;p_{0,y|x}) = \frac{\nabla_\theta F_\theta(S_{\{(1,1)\}|x;\theta})}{F_\theta(S_{\{(1,1)\}|x;\theta})} \tag{3.25}$$

$$s_\theta((0,1)|x;p_{0,y|x}) = \begin{cases} \frac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)} & \theta \in \Theta_1(x;p_{0,y|x}) \\ \frac{\nabla_\theta[\eta_1(\theta;x) - \eta_2(\theta;x)]}{\eta_1(\theta;x) - \eta_2(\theta;x)} & \theta \in \Theta_2(x;p_{0,y|x}) \\ \frac{\nabla_\theta[\eta_1(\theta;x) - \eta_3(\theta;x)]}{\eta_1(\theta;x) - \eta_3(\theta;x)} & \theta \in \Theta_3(x;p_{0,y|x}) \end{cases} \tag{3.26}$$

$$s_\theta((1,0)|x;p_{0,y|x}) = \begin{cases} \frac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)} & \theta \in \Theta_1(x;p_{0,y|x}) \\ \frac{\nabla_\theta \eta_2(\theta;x)}{\eta_2(\theta;x)} & \theta \in \Theta_2(x;p_{0,y|x}) \\ \frac{\nabla_\theta \eta_3(\theta;x)}{\eta_3(\theta;x)} & \theta \in \Theta_3(x;p_{0,y|x}) \end{cases} \tag{3.27}$$

In Section 4 below we discuss how to accurately and rapidly compute the score numerically when analytic representations are not available.

3.4. *Geometry of the Pseudo-True Set*

We next discuss the topological properties of the pseudo-true set. By its definition in (3.7), Θ^* can be viewed as the $\arg\min$-set of an optimization problem indexed by p:

$$\Theta^*(p) = \underset{\theta\in\Theta}{\arg\min}\,\phi(\theta,p), \quad \text{with} \;\; \phi(\theta,p) \equiv \inf_{q\in\mathfrak{q}_\vartheta} \int_{\mathcal{Y}\times\mathcal{X}} p(y,x) \ln \frac{p_{y|x}(y|x)}{q_{y|x}(y|x)} d\zeta(y,x). \quad (3.28)$$

THEOREM 3.2: *Suppose Θ is compact, Assumption 1 holds, and there exists a neighborhood V of p_0 and $M > 0$ such that for all $p \in V$, $\sup_{\theta\in\Theta} \mathbb{E}[\|s_\theta(Y|X; p_{y|x})\|^2] \le M$. Then the mapping $p \mapsto \Theta^*(p)$ is nonempty, compact-valued, and upper hemicontinuous.*

We prove this theorem by showing that under its assumptions, $\phi(\cdot,\cdot)$ is jointly continuous in (θ, p) (Lemma B.3), which along with compactness of Θ guarantees applicability of Berge's maximum theorem. The result establishes non-emptyness of $\Theta^*(p)$ and yields a first step towards characterizing how the geometry of the pseudo-true set changes as the extent of model misspecification changes. Given an observed density p_0, let $\{p_\gamma\} \in \Delta$, $\gamma \in \Gamma \subset \mathbb{R}$, be a net with corresponding pseudo-true values $\{\theta_\gamma\}$ such that: (i) $\theta_\gamma \in \Theta^*(p_\gamma)$ for all $\gamma \in \Gamma$, (ii) $p_\gamma \to p_0$, and (iii) $\theta_\gamma \to \theta^*$ for some $\theta^* \in \Theta$. Then Theorem 3.2 yields that $\theta^* \in \Theta^*(p_0)$: the limits of such nets form elements of the pseudo true set at p_0. This property, which holds under weak conditions, rules out that $\Theta^*(p_\gamma)$ is persistently larger than $\Theta^*(p_0)$, as $p_\gamma \to p_0$. However, it does not preclude the possibility that $\Theta^*(p_\gamma)$ shrinks from a set at p_0 (e.g., an arc in the entry game example) to a smaller set or a singleton at p_γ, no matter how close p_γ is to p_0.

This possibility is instead precluded when $p_\gamma \mapsto \Theta^*(p_\gamma)$ is lower hemicontinuous, yielding the second step characterizing how the geometry of $\Theta^*(p_\gamma)$ depends on the extent of misspecification. It is well known that lower hemicontinuity at some $\gamma_0 \in \Gamma$ is guaranteed under stronger and more challenging to verify conditions than upper hemicontinuity (Rockafellar and Wets, 2005, p.155). Equivalent statements include that for all $\theta \in \Theta$, the mapping $\gamma \mapsto \mathrm{dist}(\theta, \Theta^*(p_\gamma))$ is upper semicontinuous at γ_0; or that for every $\rho > 0$ and $\epsilon > 0$, there is a neighborhood V of γ_0 such that $\Theta^*(p_{\gamma_0})) \cap \rho\mathbb{B} \subset \Theta^*(p_\gamma)) + \epsilon\mathbb{B}$ for all $\gamma \in V \cap \Gamma$, with $\mathbb{B}$ the unit ball in $\mathbb{R}^{d_\theta}$ (e.g., Rockafellar and Wets, 2005, Propositions 5.11 and 5.12). Sufficient conditions for lower hemicontinuity are given, e.g., in Rockafellar and Wets (2005, Chapter 9) and Aubin and Frankowska (2009, Theorem 1.5.5).

Next, we provide an entry game example where we can transparently show that the correspondence $\gamma \mapsto \Theta^*(p_\gamma)$ is both lower and upper hemicontinuous, guaranteeing that the geometric properties under correct specification of the sharp identification region are preserved as the model becomes misspecified. In particular, $\Theta^*(\gamma)$ does not abruptly become a singleton as p_γ turns incompatible with the model (see Definition 3.1).

Example 1 (Specialized). Let $(U_1, U_2) \sim \mathrm{Uniform}[0,1]^2$, $\mathcal{Y} = \{(1,1),(0,1),(1,0)\}$. For $X^* \in \{0,1\}$ and $P_0(X^* = 1) = 1/2$, set $\pi_j = Y_j(\delta_{j,0}(X^*)Y_{3-j} + U_j)$, $\delta_{j,0}(X^*) = -0.5 + \theta_{j,0}X^*$, $\Theta = [-0.45, 0]^2$, and $\theta_{j,0} < 0$ for $j = 1,2$. Let $y = (0,1)$ be always selected in the region of multiplicity. Misspecification occurs because X^* is replaced by a binary proxy X with $P_0(X^* = 1|X = x) = \kappa(x,\gamma) \equiv (1-\gamma)x + \gamma(1-x)$. Expressing the observed density of $Y|X$, indexed by γ and denoted $p_\gamma(Y|X)$, as a mixture of the density $p_0(Y|X^* = x)$, for $x = 0,1$, we obtain $p_\gamma(y|X = x) = p_0(y|X^* = 1)\kappa(x,\gamma) + 0.25(1 - \kappa(x,\gamma))$, for $y = (1,1),(1,0)$, and $p_\gamma((0,1)|X = x) = 1 - p_\gamma((1,1)|X = x) - p_\gamma((1,0)|X = x)$. At $\gamma = 0$, $p_0(Y|X = x) = p_0(Y|X^* = x)$ for all $x \in \{0,1\}$ and the model is correctly specified. Regardless of the value of γ, denoting $\delta_{j,\theta}(x) = -0.5 + \theta_j x$ for some $\theta_j \in \Theta$, we have

$$\begin{aligned}\mathfrak{q}_{\theta,x} = \big\{q \in \Delta : q((1,1)|x) = (1 + \delta_{1,\theta}(x))(1 + \delta_{2,\theta}(x)),\\ -\delta_{2,\theta}(x)(1 + \delta_{1,\theta}(x)) \le q((1,0)|X = x) \le -\delta_{2,\theta}(x)\big\}.\end{aligned}$$

Ignoring possible misspecification, one would state θ's sharp identification region as $\Theta_I(p_\gamma) = \{\theta \in \Theta : p_\gamma(y|x) \in \mathfrak{q}_{\theta,y|x}, x = 0,1\} = \big\{(\vartheta - [0.5\ 0.5]^\top) \in \Theta : \vartheta_1\vartheta_2 = p_\gamma((1,1)|1);$ $\vartheta_1 \le 1 - p_\gamma((0,1)|1);$ $\vartheta_2 \le 1 - p_\gamma((1,0)|1);$ $p_\gamma((1,1)|0) = 0.25;$ $p_\gamma((1,0)|0) \in [0.25, 0.5]\big\}$. For $\gamma = 0$, $\Theta_I(p_0)$ is an arc defined by $p_0(y|1) \in \mathfrak{q}_{\theta,y|1}$, with $p_0(y|0) \in \mathfrak{q}_{\theta,y|0}$ restricting $p_0(y|0)$ but not θ. But for any $\gamma > 0$, no matter how close to 0, $\Theta_I(p_\gamma) = \emptyset$, as one can verify that $p_\gamma((1,1)|0) < (1 + \delta_{1,\theta}(0))(1 + \delta_{2,\theta}(0)) = 0.25$ for all $\theta \in \Theta$.

In contrast, our pseudo-true set $\Theta^*(p_\gamma)$ remains robust to misspecification and preserves its arc shape. After plugging the uniform distribution for F_θ into (3.11)-(3.16) to define $\eta_j(\theta;x)$ and the sets $\Theta_j(x,p_\gamma)$, and into (3.18)-(3.20) to obtain $q^*_{\theta,y|x}$ and from that $s_\theta(y|x;p_{\gamma,y|x})$ for $y\in\mathcal{Y}$, one can verify that

$$\Theta^*(p_\gamma)=\Big\{\theta\in\Theta:\ \tfrac{p_\gamma((1,1)|X=1)}{p_\gamma((0,1)|X=1)+p_\gamma((1,0)|X=1)}=\tfrac{(1+\delta_{1,\theta}(1))(1+\delta_{2,\theta}(1))}{1-(1+\delta_{1,\theta}(1))(1+\delta_{2,\theta}(1))},$$
$$\tfrac{-\delta_{2,\theta}(1)(1+\delta_{1,\theta}(1))}{1-(1+\delta_{1,\theta}(1))(1+\delta_{2,\theta}(1))}\leq\tfrac{p_\gamma((1,0)|X=1)}{p_\gamma((0,1)|X=1)+p_\gamma((1,0)|X=1)}\leq\tfrac{-\delta_{2,\theta}(1)}{1-(1+\delta_{1,\theta}(1))(1+\delta_{2,\theta}(1))}\Big\}. \quad (3.29)$$

Even when $\gamma>0$, the equality restriction in (3.29) is analogous to $p_0((1,1)|x)=(1+\delta_{1,0}(x))(1+\delta_{2,0}(x))$, $x=0,1$. Along with the other two inequalities, it yields a region with the same shape as $\Theta_I(p_\gamma)$ for $\gamma=0$. In Proposition B.2 in the Online Appendix we derive (3.29) and prove that the mapping $\gamma\mapsto\Theta^*(p_\gamma)$ is both upper and lower hemicontinuous. $\square$

3.5. *Asymptotic Distribution of the Average Score and Rao's Test Statistic*

Any $\theta^*\in\Theta^*(p_0)$ satisfies the population first order condition in (3.22). By the sample analog principle, we propose to estimate $\mathbb{E}[s_{\theta^*}(Y|X;p_{0,y|x})]$ through $\bar{s}_{\theta^*,n}(\hat{p}_{n,y|x})$, with

$$\bar{s}_{\theta,n}(p)\equiv\frac{1}{n}\sum_{i=1}^{n}s_\theta(Y_i|X_i;p),\quad \theta\in\Theta,\ p\in\Delta \quad (3.30)$$

and $\hat{p}_{n,y|x}$ a nonparametric estimator of $p_{0,y|x}$, e.g., a cell mean estimator when X has a discrete distribution or a sieve estimator when X has a continuous distribution. Our core result consists of showing that $\sqrt{n}\bar{s}_{\theta^*}(\hat{p}_{n,y|x})$ has an asymptotically normal distribution, which is insensitive to estimation of $p_{0,y|x}$. We do so leveraging the literature on semiparametric estimation, in particular Newey (1994, Proposition 2), to prove that $\mathbb{E}[s_{\theta^*}(Y,X;p_{y|x})]$ has an orthogonality property with respect to $p_{0,y|x}$. Here we provide high-level conditions under which our results attain. In Online Appendix B.2 we verify these conditions for the entry game example, both with discrete and continuous covariates. To state these conditions, for any $\theta\in\Theta$, let $m_\theta(x;p_{y|x})\equiv\mathbb{E}[s_\theta(Y|X;p_{y|x})|X=x]$. Let $\mathcal{H}$ be a parameter space to which $p_{0,y|x}$ belongs, with $\dim(\mathcal{H})=d_Y\times d_X<\infty$ if X is finitely supported, and $\mathcal{H}$ infinite dimensional otherwise. Let $\|p-p'\|_{\mathcal{H}}$ be a pseudo-metric on $\mathcal{H}$ (e.g., the sup-norm $\|p\|_{\mathcal{H}}=\sup_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}}|p(y|x)|$). For any $p\in\mathcal{H}$ and $\theta\in\Theta$, let $\mathbb{G}_{n,\theta}(p)\equiv\sqrt{n}(\bar{s}_{\theta,n}(p)-\mathbb{E}[s_\theta(Y_i|X_i;p)])$.

ASSUMPTION 2: *For each $\theta^*\in\Theta^*(p_0)$, the pathwise derivative*

$$D(\theta^*,p_{0,y|x})[p_{y|x}-p_{0,y|x}]=\lim_{\tau\to 0}\frac{\mathbb{E}[m_{\theta^*}(x,p_{0,y|x}+\tau(p_{y|x}-p_{0,y|x}))-m_{\theta^*}(x,p_{0,y|x})]}{\tau}$$

exists in all directions $(p_{y|x}-p_{0,y|x})\in\mathcal{H}$. For any $\delta_n=o(1)$ and all $\|p_{y|x}-p_{0,y|x}\|_{\mathcal{H}}\leq\delta_n$,

$$\left\|\mathbb{E}[m_{\theta^*}(X;p_{y|x})]-\mathbb{E}[m_{\theta^*}(X,p_{0,y|x})]-D(\theta^*,p_{0,y|x})[p_{y|x}-p_{0,y|x}]\right\|\leq c\|p_{y|x}-p_{0,y|x}\|^2_{\mathcal{H}}.$$

ASSUMPTION 3: *(i) The data is a random sample $(Y_i,X_i)_{i=1}^n$ drawn from P_0.*

(ii) $\hat{p}_{n,y|x}\in\mathcal{H}$ with probability approaching 1 and $\|\hat{p}_{n,y|x}-p_{0,y|x}\|_{\mathcal{H}}=o_P(n^{-1/4})$.

(iii) For each $\theta^*\in\Theta^*(p_0)$, $\mathbb{G}_{n,\theta^*}(p_{0,y|x})\xrightarrow{d}N(0,\Sigma_{\theta^*})$, with $\Sigma_{\theta^*}\equiv\mathbb{E}[s_{\theta^*}(Y_i|X_i;p_0)s_{\theta^*}(Y_i|X_i;p_0)^\top]$ the population variance-covariance matrix of the score function.

(iv) For each $\theta^*\in\Theta^*(p_0)$ and all sequences of positive numbers $\{\delta_n\}$ with $\delta_n=o(1)$,

$$\sup_{\|p_{y|x}-p_{0,y|x}\|_{\mathcal{H}}\leq\delta_n}\left\|\mathbb{G}_{n,\theta^*}(p_{y|x})-\mathbb{G}_{n,\theta^*}(p_{0,y|x})\right\|=o_P(1).$$

Assumptions 2 and 3, in their use of $\|\cdot\|_{\mathcal{H}}$, refer to the same norm. In Assumption 2, we follow Chen et al. (2003, Conditions 2.3 and 2.6) and impose a smoothness condition with respect to $p_{y|x}$ on $\mathbb{E}[s_{\theta^*}(Y,X;p_{y|x})]$. Assumption 3 (i) is a standard random sampling condition (see Epstein et al. (2016) for a discussion of inference under different assumptions). Assumption 3 (ii) requires that the estimation error of the nuisance parameter $p_{0,y|x}$ vanishes fast enough. Assumption 3 (iii) follows from the central limit theorem. Assumption 3 (iv) is a stochastic equicontinuity condition with well known primitive conditions (van der Vaart and Wellner, 1996). In Propositions B.3-B.4 in the Online Appendix, we verify all these assumptions in the two players entry game Example 1, both with discrete and continuous covariates. Under these assumptions, we obtain:

THEOREM 3.3: *Suppose Assumptions 1, 2, and 3 hold. Then, for each $\theta^* \in \Theta^*(p_0)$,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\theta^*}(Y_i|X_i;p_{0,y|x}) + o_p(1) \xrightarrow{d} N(0,\Sigma_{\theta^*}). \tag{3.31}$$

Armed with the result in Theorem 3.3, we propose to use a Rao's score statistic to test at prespecified asymptotic level $\alpha \in (0,1)$ hypotheses of the form

$$\mathbb{H}_0 : \theta^* \in \Theta^*(p_0) \text{ against } \mathbb{H}_A : \theta^* \notin \Theta^*(p_0), \tag{3.32}$$

and to obtain confidence sets by test inversion. The Rao-type test statistic takes the form

$$T_n(\theta^*) \equiv \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x})\right)^\top \tilde{\Sigma}_{n,\theta^*}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x})\right). \tag{3.33}$$

Given the score function, the test statistic in (3.33) is easy to compute even when the covariates have a continuous distribution. The weight matrix $\tilde{\Sigma}_{n,\theta^*}$ is a consistent estimator of Σ_{θ^*} when Σ_{θ^*} is not nearly singular, and assures an asymptotically valid test when Σ_{θ^*} is nearly singular, as shown below. Let $\hat{\Sigma}_{n,\theta} = \frac{1}{n}\sum_{i=1}^{n}(s_\theta(Y_i|X_i;\hat{p}_{n,y|x}) - \bar{s}_\theta(\hat{p}_{n,y|x}))(s_\theta(Y_i|X_i;\hat{p}_{n,y|x}) - \bar{s}_\theta(\hat{p}_{n,y|x}))^\top$ be the sample analog estimator of Σ_θ; $\hat{\Xi}_{n,\theta} = \hat{\Psi}_{n,\theta}^{-1/2}\hat{\Sigma}_{n,\theta}\hat{\Psi}_{n,\theta}^{-1/2}$ the correlation matrix associated with $\hat{\Sigma}_{n,\theta}$; $\hat{\Psi}_{n,\theta} = \mathrm{diag}(\hat{\Sigma}_{n,\theta})$; and $\varepsilon > 0$ a regularization constant. We recommend using the estimator proposed in Andrews and Barwick (2012, p. 2808), which introduces an adjustment insuring that the weight matrix is always nonsingular and equivariant to scale changes in the score function:

$$\tilde{\Sigma}_{n,\theta} = \hat{\Sigma}_{n,\theta} + \max\{\varepsilon - \det(\hat{\Xi}_{n,\theta}), 0\}\hat{\Psi}_{n,\theta}, \quad \theta \in \Theta. \tag{3.34}$$

COROLLARY 3.1: *Let Assumptions 1, 2, and 3 hold. Then, under $\mathbb{H}_0$ in (3.32), for any $\theta^* \in \Theta^*(p_0)$ such that $\min_{j=1,\ldots,d_\theta}\{\mathrm{diag}(\Sigma_{\theta^*})\}_j > 0$ and*

$$\hat{\Sigma}_{n,\theta^*} \xrightarrow{p} \Sigma_{\theta^*}, \tag{3.35}$$

(a) If Σ_{θ^*} is nonsingular, $T_n(\theta^*) \xrightarrow{d} \chi^2_{d_\theta}$; (b) Both for singular and nonsingular Σ_{θ^*}, $\limsup_{n\to\infty} P(T_n(\theta^*) > c_{d_\theta,\alpha}) \leq \alpha$, with $c_{d_\theta,\alpha}$ the $1-\alpha$ quantile of the $\chi^2_{d_\theta}$ distribution.

Example 1 (Continued). The score function for Example 1 on p. 17 implies that $\Theta^*(p_\gamma)$ is characterized by a single equality restriction, and hence the rank of Σ_{θ^*} in this case is lower than d_θ and the critical value in Corollary 3.1 is asymptotically conservative. □

Corollary 3.1 requires, in (3.35), that the population covariance matrix can be consistently estimated. In the semiparametric literature with point identification, this is a standard requirement;[9] in the moment inequalities literature with partial identification, it is also common to assume that the covariance matrix of the moment functions can be consistently estimated for all θ in the identified set. The result in Corollary 3.1 is valuable because it implies that no simulations are needed to compute the quantiles of the limiting distribution, and that the critical values used to test the hypothesis in (3.32) and to construct the confidence set via test inversion are constant across candidates $\theta \in \Theta$. This is in contrast with much of the related literature, where the asymptotic distribution of the test statistic is nonpivotal and the critical values need to be recomputed for each θ.[10] One can construct a confidence region that covers each point in $\Theta^*(p_0)$ with asymptotic probability $1-\alpha$ as

$$CS_n = \left\{\theta \in \Theta : T_n(\theta) \leq c_{d_\theta,\alpha}\right\}. \tag{3.36}$$

In practice, CS_n is computed by specifying a grid of values Θ_n through which to explore the parameter space, and letting $CS_n = \{\theta \in \Theta_n : T_n(\theta) \leq c_{d_\theta,\alpha}\}$.

We next show that CS_n is an asymptotically *uniformly valid* confidence set. We posit that P_0, the distribution of the observed data, belongs to a class of distributions denoted by $\mathcal{P}$, where the conditional law $P(\cdot|x)$ for each $P \in \mathcal{P}$ is absolutely continuous with respect

[9] In the semiparametric literature, (3.35) is imposed for $\hat{\Sigma}_{n,\hat{\theta}_n}$, with $\hat{\theta}_n$ a consistent estimator of a singleton θ^*; in our use of it, $\hat{\Sigma}_{n,\theta^*}, \Sigma_{\theta^*}$ are both evaluated at the same θ^*, but the requirement applies to all $\theta^* \in \Theta^*(p_0)$.

[10] Nonpivotal asymptotic distributions appear, e.g., in Andrews and Kwon (2024), Andrews and Shi (2013), Kaido et al. (2019), and Bugni et al. (2017), while Chen et al. (2018)'s test statistic converges to χ^2 distribution.

to μ on $\mathcal{Y}$. We let $p_{y|x}$ denote the Radon-Nykodim derivative of $P(\cdot|x)$. We write stochastic order relations that hold uniformly over $P \in \mathcal{P}$ using the notations $o_{\mathcal{P}}$ and $O_{\mathcal{P}}$.

THEOREM 3.4: *For constants $c > 0$ and all $P \in \mathcal{P}$, let Assumptions 1, 2, and 3 hold, with the following conditions replacing the corresponding ones in the original assumptions:*[11]

1(b)' $\mathcal{A}_G = \operatorname{supp}(G(\cdot|X;\theta)) \equiv \{A \subseteq \mathcal{Y} : F_\theta(G(U|X;\theta) = A) > c\}$ *for all* $\theta \in \Theta$, $P - a.s.$

1(d)' $\Theta^*(p) \subset \operatorname{int} \Theta^{-c} \equiv \{\theta \in \Theta : B_c(\theta) \subset \Theta\}$.

1(e)' *For all* $\theta \in \Theta$ *and* $y \in \mathcal{Y}$, $q^*_{\theta,y|x}(y|x) > c$, $P - a.s.$

2' *The constant* c *is the same for all* $P \in \mathcal{P}$.

3(ii)' *For all* $\epsilon > 0$ *there exists* $N \in \mathbb{N}$, *with* ϵ *and* N *not dependent on* $P \in \mathcal{P}$, *such that* $P(\hat{p}_{n,y|x} \in \mathcal{H}) \geq 1 - \epsilon$, $\forall n \geq N$, *and* $\|\hat{p}_{n,y|x} - p_{0,y|x}\|_{\mathcal{H}} = o_{\mathcal{P}}(n^{-1/4})$.

3(iv)' *For all sequences of positive numbers* $\{\delta_n\}$ *with* $\delta_n = o(1)$,

$$\sup_{\theta^* \in \Theta^*(p_0)} \sup_{\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta_n} \left\| \mathbb{G}_{n,\theta^*}(p_{y|x}) - \mathbb{G}_{n,\theta^*}(p_{0,y|x}) \right\| = o_{\mathcal{P}}(1).$$

Suppose that for all $P \in \mathcal{P}$, $\min_{j=1,\dots,d_\theta}\{\operatorname{diag}(\Sigma_{\theta^*})\}_{\mathrm{j}} > 0$, *and* $\|\hat{\Sigma}_{n,\theta^*} - \Sigma_{\theta^*}\| = o_{\mathcal{P}}(1)$. *Then, for* CS_n *in* (3.36), *we have*

$$\liminf_{n\to\infty} \inf_{P \in \mathcal{P}} \inf_{\theta^* \in \Theta^*(p)} P(\theta^* \in CS_n) \geq 1 - \alpha.$$

Under the assumptions of Theorem 3.4, Corollary 3.1 also applies uniformly over $P \in \mathcal{P}$.

REMARK 3.2: *Our confidence set can be interpreted as the union over $\theta^* \in \Theta^*$ of ellipsoids centered at θ^* whose volume is determined by $\tilde{\Sigma}_{n,\theta^*}$ only, and not by the extent of misspecification. The lever through which the latter might impact the volume of CS_n, is through Θ^*'s volume, which may or may not be impacted by misspecification (see Section 3.4). Relative to Θ^*'s volume, CS_n's volume depends only on the sampling variability of the score statistic, and as Θ^* is always non-empty, CS_n is always non-empty as well.*

[11]The constants c may differ across appearances but do not depend on P; $\mathbb{N}$ denotes the natural numbers; and $B_c(\theta)$ denotes a ball of radius c centered at θ.

3.6. *Role of Sharp Identifying Restrictions*

Our analysis is able to exploit, through (3.2), all identifying restrictions associated with the economic model. Doing so is in part motivated by our desire to avoid the risk of potentially discordant conclusions driven by model misspecification and different choices of subsets of moment inequalities on which to base inference (an issue well articulated in Li et al., 2024). And in part by the fact that it allows us to provide a general proof, under Assumption 1-(b), of Lemma B.2. This lemma establishes the existence of a representation of $L(\theta|x)$, for all $\theta \in \Theta$, characterized by $q^*_{\theta,y|x}$ and unique Lagrange multipliers. Importantly, one does not need to calculate that representation, because $L(\theta|x)$ coincides with it pointwise (see Lemma B.2-(ii)). Yet, $L(\theta|x)$ in (3.10) inherits its differentiability properties.[12]

In practice, when the number of inequalities in (3.2) is relatively small, users of our method can derive the score in closed form. When it is moderate (a few hundred), the numerical score can be computed directly based on (3.2) or many redundant inequalities can be quickly eliminated without resorting to specialized methods.[13] When the number of inequalities in (3.2) is substantially larger, one can use Algorithm 3 in Luo et al. (2025) to eliminate all redundant ones. This algorithm remains practically feasible, with reported computational times of seconds for a Julia implementation that removes all redundant inequalities in entry game examples where (3.2) includes 10^{14} inequalities, and dynamic discrete choice examples with 10^{154} inequalities (Luo et al., 2025, Table 1 and p.23).

Yet, when the number of inequalities is prohibitive, one may need to carry out inference based on a subset of inequalities.[14] Then, our procedure remains valid provided the conclusions of Lemma B.2 continue to hold. Hence, among the criteria guiding the selection of inequalities, one might include ensuring this is the case. We note that a rich literature studies how to remove redundant constraints in linear programs. For a given chosen subset

[12]The presence of redundant (in)equality constraints in the original formulation of $L(\theta|x)$ does not affect differentiability, as there exists an equivalent formulation without such redundancies with unique Lagrange multipliers.

[13]See the Python library created by Hiroaki Kaido to carry out this task for discrete choice models, and the computational simplifications in Bontemps and Kumar (2020) for a certain class of multi-players entry models.

[14]Inference methods that rely on discretization of $\mathcal{X}$ and aim to use all information in the model may face an additional computational bottleneck, as those methods have a final number of inequalities given at least by the number of nonredundant inequalities in (3.2) multiplied by the cardinality of the discretization of $\mathcal{X}$.

of inequalities in (3.2), which by construction defines a set linear in q that includes $\mathfrak{q}_\theta$, as long as the identity of the non-redundant inequalities does not change with θ, the conclusions of Lemma B.2 continue to hold (for the new linear program with enlarged constrained set) and our method remains valid. In this case, one obtains a different pseudo-true set than the one yielded by sharp restrictions, with weakly lower minimal KL divergence, and our confidence set then covers the elements of this different pseudo-true set with a prespecified asymptotic probability. Without the conclusions of Lemma B.2, existence of the score is no longer guaranteed, rendering a score-based approach inapplicable. A likelihood-based approach may nonetheless be possible, but is beyond the scope of this paper.

Example 1 (Non-sharp). Consider the two player entry game on p. 6 with $\mathfrak{q}_\theta$ in (3.6) and $\delta_1 < 0, \delta_2 < 0$. Instead of using sharp identifying restrictions, replace $\mathfrak{q}_\theta$ with $\{q_{y|x} \in \Delta: \ q_{y|x}(y|x) \geq F_\theta(S_{\{y\}|x;\theta}), \ y \in \mathcal{Y}, x \in \mathcal{X}\}$. Then $L(\theta|x)$ in (3.10) can be represented using a set of constraints of the form $Aq_{y|x} \geq b(\theta)$, with A a 4×4 matrix with three rows equal to standard basis vectors and one a vector of 1. Hence, Lemma B.2 holds. □

4. COMPUTATION OF THE SCORE FUNCTION

Sometimes it is possible to obtain a closed-form expression for $s_\theta(y|x;p_{0,y|x})$ as gradient of $\ln q^*_{\theta,y|x}$ with respect to θ, as in Example 1 (p. 15). If $q^*_{\theta,y|x}$ does not have a closed form expression, one needs to compute the score numerically. Here we describe how to do so, adapting the method in Forneron (2023). We omit the dependence of $s_\theta(y|x)$ on $p_{0,y|x}$ or its estimator. We presume that one can compute $q^*_{\theta,y|x}$ relatively easily (e.g., using `cvxpy`).

Consider a smoothed version f_ς of $f(\theta) \equiv \ln q^*_{\theta,y|x}$, defined by the convolution:

$$f_\varsigma(\theta) = \int f(\theta + \varsigma z)\phi(z)dz,$$

where ϕ is a smooth kernel decaying to 0 in the tails, such as the Gaussian density function. The derivative of f_ς exists. If it admits integration by parts, one has:

$$\tfrac{\partial}{\partial\theta} f_\varsigma(\theta) = -\tfrac{1}{\varsigma}\int f(\theta+\varsigma z)\tfrac{\partial}{\partial z}\phi(z)dz = -\tfrac{1}{\varsigma}\mathbb{E}\Big[f(\theta+\varsigma Z)\frac{\frac{\partial}{\partial z}\phi(Z)}{\phi(Z)}\Big],$$

Algorithm 1 Construct CS_n

Data: $W^n = (Y_i, X_i), i = 1, \dots, n$

Require: $\varepsilon > 0$, $K_n > 0$, $\varsigma > 0$ [if no closed form for score function, else $\varsigma = 0$], Θ_n

$\hat{p}_{n,y|x} \leftarrow$ CCP(W^n;K_n) [K_n the tuning parameter for $\hat{p}_{n,y|x}$]

$\bar{s}_{\theta;\varsigma}(Y_i|X_i;\hat{p}_{n,y|x}) \leftarrow$ SCORE(W^n;$\hat{p}_{n,y|x}$;θ;ς) [(4.2) if no closed form, else (3.30)]

$c_{d_\theta,\alpha} \leftarrow \inf\{x : P(\chi^2_{d_\theta} \geq x) \leq \alpha\}$

$CS_n \leftarrow \emptyset$

for $\theta \in \Theta_n$ **do**

  $\tilde{\Sigma}_{n,\theta} \leftarrow$ REGCOVMAT(W^n,θ;ε) [(3.34)]

  $T_n(\theta) \leftarrow \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n s_\theta(Y_i|X_i;\hat{p}_{n,y|x})\right)^\top \tilde{\Sigma}^{-1}_{n,\theta} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n s_\theta(Y_i|X_i;\hat{p}_{n,y|x})\right)$

  if $T_n(\theta) \leq c_{d_\theta,\alpha}$ **then**

    Add θ to CS_n

  end if

end for

return CS_n

with the last expectation taken with respect to $Z \sim N(0, I_d)$. One can then approximate the derivative of f by that of f_ς. Letting $Z_r, r = 1, \dots, R$ be i.i.d. draws from $N(0, I_d)$, and noting that $\frac{\frac{\partial}{\partial z}\phi(Z)}{\phi(Z)} = \nabla \ln \phi(z) = -z$, an unbiased estimator for $\frac{\partial}{\partial \theta} f_\varsigma(\theta)$ is

$$\frac{1}{\varsigma R} \sum_{r=1}^R [f(\theta + \varsigma Z_r) - f(\theta)] Z_r. \tag{4.1}$$

Replacing f with f_ς introduces a bias proportional to ς, while the variance of $[f(\theta + \varsigma Z_r) - f(\theta)] Z_r / \varsigma$ grows with $1/\varsigma^2$. In practice one needs to take a stand on this bias-variance trade off. We formally analyze how to do so in research-in-progress (available from the authors upon request), where numerical experiments lead us to recommend setting $\varsigma = c(nR)^{-1/4}$ for $c \in [0.03, 0.12]$. The Monte Carlo approximation in (4.1) inflates the variance as well, by a factor of $(1 + R^{-1})$ as in the method of simulated moments. This factor can easily be incorporated in the estimator of the asymptotic variance of the score. Letting $f(\theta; Y_i, X_i) = \ln q^*_{\theta,y|x}(Y_i, X_i)$, one can obtain the estimator in (4.1) for each value of (Y_i, X_i). The average score can then be approximated by

$$\bar{s}_{\theta;\varsigma}(\hat{p}_{n,y|x}) = \frac{1}{n\varsigma R} \sum_{i=1}^n \sum_{r=1}^R [f(\theta + \varsigma Z_{i,r}; Y_i, X_i) - f(\theta; Y_i, X_i)] Z_{i,r}. \tag{4.2}$$

Algorithm 1 presents pseudo-code with the steps and tuning parameters required to build the confidence set in (3.36). When a closed form expression for $s_\theta(Y_i|X_i;\hat{p}_{n,y|x})$ is available, one can set $\varsigma=0$ and plug the observed values $\{Y_i,X_i\}_{i=1}^n$ in (3.30); when it is not available, one needs to choose $\varsigma>0$ and plug $\{Y_i,X_i\}_{i=1}^n$ in (4.2).

5. EMPIRICAL ILLUSTRATION

We illustrate the usefulness of our method by applying it to answer the question addressed in Kline and Tamer (2016, Section 8): "what explains the decision of an airline to provide service between two airports." Kline and Tamer analyze data for the second quarter of the year 2010, documenting the entry decisions of two types of airline companies: Low Cost Carriers (LCC) versus Other Airlines (OA).[15] They define a market as a trip between two airports, irrespective of intermediate stops. They record the entry decision $Y_{j,m}$ of player $j\in\{LCC,OA\}$ in market m as a 1 if a firm of type j serves market m, and 0 otherwise. They posit that player j's decision to serve a market depends not only on their opponent's entry decision, but also on observable payoff shifters $X_{j,m}$ and unobservable payoff shifters $U_{j,m}$. The observable payoff shifters $X_{j,m}$ include the constant and two continuously distributed variables, $X_{j,m}^{pres}$ and X_m^{size}. The first variable is firm-and-market-specific: it measures the market presence of firms of type j in market m (see Kline and Tamer, 2016, p. 356 for its exact definition). Market presence of the LCC airline, $X_{LCC,m}^{pres}$ (respectively, $X_{OA,m}^{pres}$), is excluded from the payoff of firm OA (respectively, LCC). The second variable, market size, enters the payoff of firms of both types; it measures population size at the two endpoints of the trip and is market-specific. The unobservables $U_{j,m}$, $j\in\{LCC,OA\}$, are assumed to have a bivariate normal distribution with $\mathbb{E}(U_{j,m})=0$, $Var(U_{j,m})=1$, $Corr(U_{LCC,m},U_{OA,m})=r$, and to be i.i.d. across m.[16]

Both Kline and Tamer (2016) and we assume that players enter the market if doing so yields non-negative payoffs. However, we posit different payoff functions. They posit:

[15]We use their data, downloading it from Quantitative Economics' online repository.

[16]We assume $r\in[-0.9,0.9]$ and estimate it as part of the vector θ. We ensure that the strategic interaction parameters δ_{LCC} and δ_{OA} are less than a constant $c<0$ and that $q_{\theta^*,y|x}>c$ for another constant $c>0$.

$$\tilde{\pi}_{j,m} = Y_{j,m}(\tilde{\beta}_j^0 + \tilde{\beta}_j^{pres}\mathbf{1}(X_{j,m}^{pres} \geq Med(X_j^{pres})) \\ + \tilde{\beta}_j^{size}\mathbf{1}(X_m^{size} \geq Med(X^{size})) + \tilde{\delta}_j Y_{-j,m} + U_{j,m}). \quad (5.1)$$

In words, Kline and Tamer transform each of market size and of the two market presence variables into binary variables, based on whether each of these variables realizes above or below their respective median. Doing so yields a finite number of unconditional moment inequalities, which they need for their inference procedure, at the cost of foregoing the information provided in the variation in $X_{j,m}$ past whether each variable is above or below its median, and of using an arguably more restrictive payoff function.

Leveraging our new method, we are able to avoid discretizing the continuously distributed covariates, thereby exploiting all identifying power in their variation and allowing $X_{j,m}$ to impact payoffs proportionally to their value. We assume that payoffs take the form:

$$\pi_{j,m} = Y_{j,m}(\beta_j^0 + \beta_j^{pres}X_{j,m}^{pres} + \beta_j^{size}X_m^{size} + \delta_j Y_{-j,m} + U_{j,m}). \quad (5.2)$$

We study how the decision of an LCC airline to enter the market is affected by whether an OA airline is in the market and by the extent of LCC airlines market presence. To do so, we define the potential entry decision of an LCC player as $Y_{LCC}(d) = \mathbf{1}(X_{LCC,m}^\top\beta_{LCC} + \delta_{LCC}d + U_{LCC,m} \geq 0)$, with $\beta_j = (\beta_j^0, \beta_j^{pres}, \beta_j^{size})$, $j \in \{LCC, OA\}$. This is the entry outcome of an LCC airline when we fix the OA's entry to take value $d \in \{0,1\}$. Based on our model, the entry probability of the LCC airline is $P(Y_{LCC}(d) = 1|X_{LCC,m}) = \Phi(X_{LCC,m}^\top\beta_{LCC} + \delta_{LCC}d)$. We obtain a confidence interval for this parameter for each of $d = 0,1$ and for specific values of $X_{LCC,m}$. To ease the reporting of results, we set X_m^{size} equal to the median of its distribution and compute the τ-quantile of the distribution of $X_{LCC,m}^{pres}$ for $\tau \in \mathcal{T} \equiv \{0.125, 0.250, 0.375, 0.5, 0.625, 0.750, 0.875\}$; we evaluate our parameter of interest for $X_{LCC,m}^{pres}$ set equal to each of these values, but our confidence set construction uses all information in $X_{j,m}$. Letting $\theta = (\beta_{LCC}, \delta_{LCC}, \beta_{OA}, \delta_{OA}, r)$, we report confidence intervals $CI_n(x,d) = \big[\min_{\theta:\, T_n(\theta) \leq c_{d_\theta,\alpha}} \Phi(x_{LCC,m}^\top\beta_{LCC} + \delta_{LCC}d)$, $\max_{\theta:\, T_n(\theta) \leq c_{d_\theta,\alpha}} \Phi(x_{LCC,m}^\top\beta_{LCC} + \delta_{LCC}d)\big]$ for the values of x corresponding to $\tau \in \mathcal{T}$

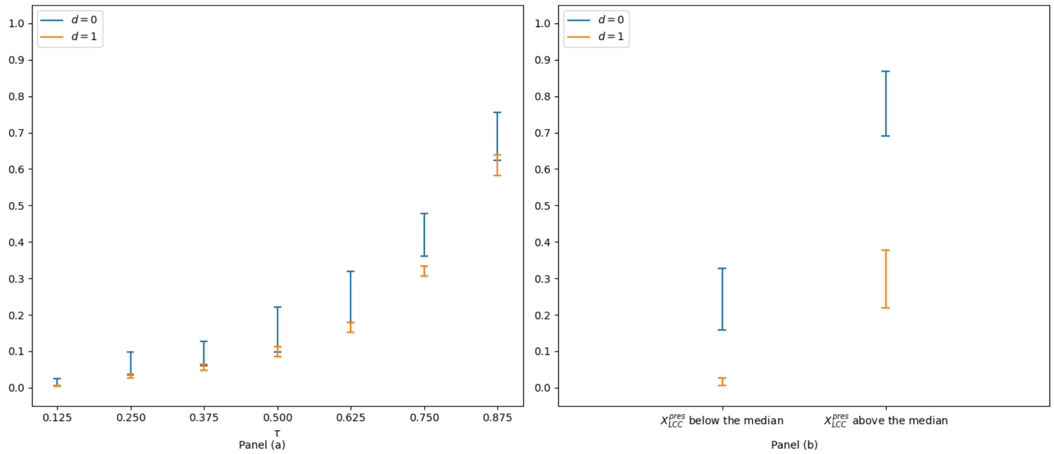

FIGURE 2.—Confidence Intervals for $\Phi(x_{LCC,m}^{\top}\beta_{LCC}^{*}+\delta_{LCC}^{*}d)$ for $d=1$ (orange) and $d=0$ (blue). Panel (a): Rao score test-based inference with $X_{LCC,m}^{pres}$ set equal to the τ quantile of its distribution and X_{m}^{size} set equal to its median. Panel (b): Chen et al. (2018) projection-based inference with $\mathbf{1}(X_{j,m}^{pres} > Med(X_{j}^{pres}))$.

and for $d \in \{0,1\}$.[17] Under the conditions of Theorem 3.4, by standard arguments

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta^{*}\in\Theta^{*}(p)} P(\Phi(x_{LCC,m}^{\top}\beta_{LCC}^{*}+\delta_{LCC}^{*}d) \in CI_{n}) \geq 1-\alpha.$$

Figure 2-Panel (a) reports our results, displaying on the horizontal axis the value of τ and on the vertical axis the candidate value for $\Phi(x_{LCC,m}^{\top}\beta_{LCC}^{*}+\delta_{LCC}^{*}d)$. The results show potentially sizable heterogeneity in the treatment effects of interest and reject the hypothesis that they are constant across τ. While the confidence intervals for $d=0,1$ are not disjoint for several values of τ, when OA opponents are not in the market (blue segments in Figure 2 for $d=0$) the entry probability can be much larger than when they are present (orange segments for $d=1$) across all values of τ, with the effect largest for $\tau=0.750$.

For each fixed value of the entry decision of OA, as the market presence of the LCC airlines increases, so does the probability that LCC firms enter a market. When OA firms are in the market, the difference is statistically significant for all τ, although the impact of $x_{LCC,m}^{pres}$ on the entry probability is low until market presence reaches its 0.625 quantile,

[17]As in the Monte Carlo experiments in Section 6, we estimate $p_{0,y|x}$ using a series estimator with J-th order (tensor product) B-spline basis functions and set $\varepsilon=0.05$ in (3.34) to compute $\tilde{\Sigma}_{n,\theta}$.

at which point the slope increases rapidly. This suggests that in order to overcome the presence of OA opponents and enter the market, LCC firms need large market presence. On the other hand, when OA firms are not in the market, the impact of $x^{pres}_{LCC,m}$ on the entry probability is sizable starting with $\tau = 0.375$ (and further increases with τ).

We compare our results to what one would obtain using the likelihood based inference method in Chen et al. (2018), which is designed for correctly specified models with discrete covariates. We note that Chen et al. assume the payoff function in (5.1), whereas we use the specification in (5.2), and therefore the coefficient estimates on $X_{j,m}$ and Y_{-j} are not directly comparable to each other. Nonetheless, we believe it to be instructive to compare the counterfactual model-implied entry probabilities across the two approaches.[18] Figure 2-Panel (b) reports confidence intervals based on Chen et al. (2018)'s projection method for $d = 0, 1$ and for $X^{pres}_{LCC,m}$ below the median and above the median. The figure shows that aggregating the value of $X^{pres}_{LCC,m}$ at this coarse level hides interesting patterns in the results. Bundling "above the median" and "below the median" as single values for the covariates does not allow one to learn the extent of the heterogeneity in the effect of market presence on the probability of entry. Moreover, using Chen et al. (2018)'s inference method yields very large treatment effects for the presence of an OA opponent, which instead we do not find. Of note, our confidence sets are shorter than those based on Chen et al. (2018) for all values of τ, often substantially, and hence this difference is not driven by a reduction in precision but likely by a different model and our use of all information in the covariates. While one could use a finer discretization of $(X^{pres}_{LCC,m}, X^{pres}_{OA,m})$ in (5.1) combined with Chen et al. (2018)'s method, doing so would result in a substantially harder computational problem. Indeed, in Chen et al.'s approach the selection probabilities, which are allowed to depend on X (but not U) and have cardinality at least equal to the cardinality of $\mathcal{X}$, are part of the parameters to be estimated. In contrast, the computational complexity of our procedure does not change with the cardinality of $\mathcal{X}$.

[18] We use the replication package provided by Chen et al. (2018) at Econometrica's online repository, where the payoffs are specified as $\tilde{\pi}_{j,m} = Y_{j,m}(\tilde{\beta}^0_j + \tilde{\beta}^{pres}_j \mathbf{1}(X^{pres}_{j,m} > Med(X^{pres}_j)) + \tilde{\beta}^{size}_j \mathbf{1}(X^{size}_m > Med(X^{size})) + \tilde{\delta}_j Y_{-j,m} + U_{j,m})$ (compare with 5.1). We compute the confidence intervals using the payoffs in their code and obtain a confidence interval on $P(Y_{LCC}(d) = 1 | X_{LCC,m})$ through the projection method that they propose.

TABLE I

MAXIMUM LIKELIHOOD ESTIMATE OF (θ,κ)

β^0_{LCC}	β^{pres}_{LCC}	β^{size}_{LCC}	δ_{LCC}	β^0_{OA}	β^{pres}_{OA}	β^{size}_{OA}	δ_{OA}	κ
-0.367	2.044	-0.066	-0.085	0.282	1.774	0.251	-0.226	0.000

6. MONTE CARLO EXPERIMENTS

We carry out an empirical Monte Carlo exercise where the data-generating process is calibrated to the data we use for the application in Section 5; the notation is as in that section and the payoffs are as in (5.2). We normalize each covariate in $\{X^{pres}_{LCC}, X^{pres}_{OA}, X^{size}\}$ to the unit interval and assume $(U_{LCC,m}, U_{OA,m})$ is distributed i.i.d. bivariate standard normal. We let $\theta = (\beta_{LCC}, \delta_{LCC}, \beta_{OA}, \delta_{OA})$.

To calibrate a DGP value for the parameter vector θ from Kline and Tamer (2016)'s data, we introduce a parameter $\kappa \in [0,1]$, independent of (X,U), representing the probability of selecting outcome $Y=(1,0)$ when multiple equilibria are present. Define

$$q_{(\theta,\kappa),y|x}((0,0)|x) = [1-\Phi(x_{LCC}^\top\beta_{LCC})][1-\Phi(x_{OA}^\top\beta_{OA})] \tag{6.1}$$

$$q_{(\theta,\kappa),y|x}((1,1)|x) = \Phi(x_{LCC}^\top\beta_{LCC}+\delta_{LCC})\Phi(x_{OA}^\top\beta_{OA}+\delta_{OA}), \tag{6.2}$$

$$q_{(\theta,\kappa),y|x}((0,1)|x) = \eta_1(\theta;x) - q_{\theta,y|x}((1,0)|x) \tag{6.3}$$

$$q_{(\theta,\kappa),y|x}((1,0)|x) = \eta_3(\theta;x) + \kappa(\eta_2(\theta;x)-\eta_3(\theta;x)) \tag{6.4}$$

where the functions $\eta_1(\cdot;x), \eta_2(\cdot;x), \eta_3(\cdot;x)$ are defined in (3.11), (3.12), (3.13). We report in Table I the value for (θ,κ) estimated by maximizing the likelihood based on (6.1)-(6.4). We use this estimate as baseline DGP value in our simulations.[19]

We consider two designs for our DGPs. Design 1 uses only the two player-specific covariates, $\mathbf{X}^{D1} = (X^{pres}_{LCC}, X^{pres}_{OA})$, and sets $(\theta_0,\kappa_0) = (\beta^0_{LCC}, \beta^{pres}_{LCC}, \delta_{LCC}, \beta^0_{OA}, \beta^{pres}_{OA}, \delta_{OA}, \kappa)$

[19]To ensure that the DGP induces conditional choice probabilities that are bounded away from 0 and 1, we estimate (θ,κ) using observations such that the estimated conditional choice probabilities are in the interval $(\epsilon, 1-\epsilon)$ with $\epsilon = 1e-3$. This gives us a sample of size 7,017.

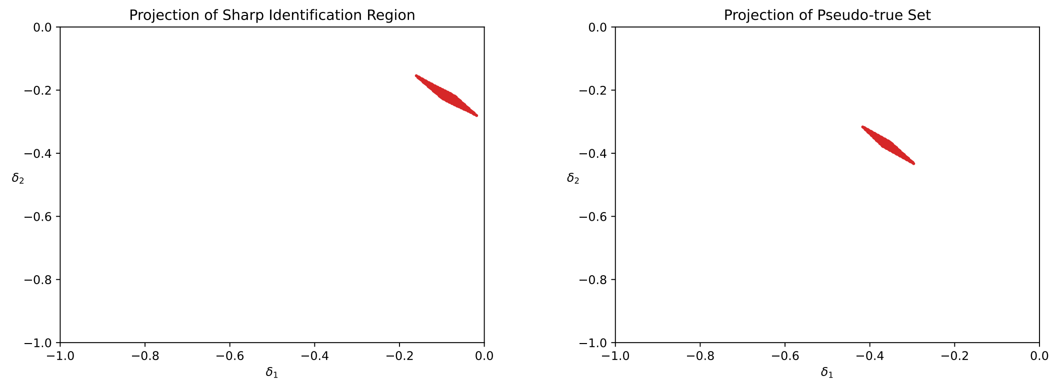

FIGURE 3.—Projections of $\Theta^*(p_0)$ (left, correctly specified; right, misspecified with $\gamma = -0.4$)

to the corresponding maximum likelihood estimates in Table I. Design 2 incorporates the full set of covariates, $\mathbf{X}^{D2} = (X^{pres}_{LCC}, X^{pres}_{OA}, X^{size})$ and sets (θ_0, κ_0) to the full MLE vector in Table I. As we further explain below, we implement Design 1 because of computational difficulties with Design 2 for the comparator inference method of Andrews and Shi (2013).

Within each design $k = 1, 2$, we resample from the original Kline and Tamer (2016)'s dataset covariates $\mathbf{X}^{Dk}_m, m = 1, \ldots, n$. We evaluate the performance of our inference method (and the Andrews and Shi, 2013, comparator method) when the model is correctly specified, by generating $(Y_{LCC,m}, Y_{OA,m})$, $m = 1, \ldots, n$ through inverse CDF transformation, using as DGP the probability mass functions in (6.1)-(6.4) in which we plug $\mathbf{X}^{Dk}_m$ and, for (θ, κ), the MLE values in Table I for that design.

To evaluate performance under model misspecification, we simulate data from a two-player entry game where the true payoff for player j in Design k is given by:

$$\pi^{Dk}_j = Y^{Dk}_j (X^{Dk\top}_j \beta^{Dk}_j + (\delta_j + \gamma X^*) Y^{Dk}_{-j} + U_j),\ Y^{Dk}_j \in \{0,1\},\ j \in \{LCC, OA\},$$

with X^* a binary variable omitted from the model, $X^{D1}_j = [1 \quad X^{pres}_j]^\top$, and $X^{D2}_j = [1 \quad X^{pres}_j \quad X^{size}]^\top$. Given $(X^{pres}_{LCC}, X^{pres}_{OA}) = (\tilde{x}_{LCC}, \tilde{x}_{OA})$, $X^* = 1$ with probability $\Phi(\frac{\tilde{x}_{LCC} - \mu_{LCC}}{\sigma_{LCC}} + \frac{\tilde{x}_{OA} - \mu_{OA}}{\sigma_{OA}})$, with μ_j and σ^2_j the mean and variance of $X^{pres}_{j,m}$. The value of γ determines the extent of misspecification. We report results for $\gamma \in \{-.1, -.2, -.3, -.4\}$.

Figure 3 shows the projections of $\Theta^*(p_0)$ onto the space of $(\delta_{LCC}, \delta_{OA})$ under Design 2.[20] When the model is correctly specified (left panel), $\Theta^*(p_0)$ coincides with the sharp identification region of θ. With misspecification ($\gamma = -0.4$), the optimal value of the KL divergence measure in (3.7) is strictly positive (see Table II for the value of $I(p_0||q^*_\theta)$), indicating that the sharp identification region is empty. In contrast, the pseudo-true set $\Theta^*(p_0)$ remains nonempty, as shown in the right panel of Figure 3, and the shape of $\Theta^*(p_0)$ remains similar both in the correctly specified and in the misspecified case. Nonetheless, under misspecification it shifts to the (lower) left. This is because when $X^* = 1$, the true DGP allocates a large mass to either $(1,0)$ or $(0,1)$ (whereas with $X^* = 0$ that mass is allocated to $(1,1)$). This is not captured by the model in (5.2). A reduction in the values of (δ_1, δ_2) enlarges the region of multiplicity, thereby allowing for a larger mass to be allocated to $(1,0)$ and $(0,1)$ than the model in (5.2) allows for.

To implement our test, we estimate $p_{0,y|x}$ by a series estimator with J-th order (tensor-product) B-spline basis functions.[21] For Design 1, we compare the performance of our procedure to that of Andrews and Shi (2013).[22] We report a power comparison where the parameter value θ^* lies on the boundary of $\Theta^*(p_0)$. Specifically, we examine the rejection probabilities for local alternatives of the form $\theta_{0,h} = \theta^* + h \times (1,1,0,\ldots,0)'/\sqrt{n}$, with $h > 0$. This corresponds to drifting the strategic interaction effects toward $(0,0)$ while holding the other components fixed. Andrews and Shi's test transforms the conditional moment inequalities into unconditional ones using as instruments indicator functions of whether each covariate belongs to specified hypercubes. The side length of each hypercube is $1/(2r)$, for $r = 1, \ldots, r_{1n}$, where a larger value of r_{1n} corresponds to finer conditioning

[20] A similar figure for Design 1 is omitted to conserve space but available from the authors upon request.

[21] We compute $\tilde{\Sigma}$ in (3.34) setting $\varepsilon = 0.05$ as done in Andrews and Shi (2013); additional simulations (available from the authors) with $\varepsilon = 0.025$ and 0.1 indicate that increasing ε makes our procedure more conservative.

[22] We implement the method in Andrews and Shi (2013) using their S_3 test statistic, with moment functions

$$m_{\leq}(Z_i,\theta) = \begin{bmatrix} 1\{Y_i = (1,0)\} - \eta_2(x;\theta) \\ \eta_3(x;\theta) - 1\{Y_i = (1,0)\} \end{bmatrix};\ m_{=}(Z_i,\theta) = \begin{bmatrix} 1\{Y_i = (0,0)\} - [1 - \Phi(x_{LCC}^\top \beta_1)][1 - \Phi(x_{OA}^\top \beta_2)] \\ 1\{Y_i = (1,1)\} - \Phi(x_{LCC}^\top \beta_1 + \delta_1)\Phi(x_{OA}^\top \beta_2 + \delta_2)\} \end{bmatrix},$$

where $m(z,\theta) = (m_{\leq}(z,\theta)^\top, m_{=}(z,\theta)^\top)^\top$, and with $\bar{m}_n(\theta) = \frac{1}{n}\sum_{i=1}^n m(Z_i,\theta)$. The number of inequalities is two times the number of hyper-cubes used, and similarly for the equalities.

information. Following Andrews and Shi (2013, Section 10.4), we set $r_{1n} = 3$, and also report results for $r_{1n} = 2$ and 4. For Design 2, we report results only for the score-based test, as the moment-based test was computationally infeasible in this setting.

Table II reports the results of this exercise for 500 Monte Carlo repetitions. Panel (A) documents the size and power of our test as well as the moment inequality-based tests for the case that the model is correctly specified. The test of Andrews and Shi (2013) over-rejects slightly in this correctly specified DGP, while our Rao's score-based test has valid size but under-rejects. Nonetheless, the power curve of our test quickly dominates that of the moment inequality based test.

In the misspecified case (Panels (B)-(F)), as expected the moment inequality-based test is oversized. The extent of the size distortion grows with the extent to which the model is misspecified. To quantify the latter across DGPs, we compute the rejection probability of an infeasible Information Matrix test (White, 1982) that uses knowledge of the fact that the selection mechanism R in (3.3) is distributed $\mathrm{Bernoulli}(\kappa_0)$. Enriched with this information, the model yields a unique prediction $q_{(\theta,\kappa),y|x}$ and a well defined likelihood function, and hence we can obtain the (point identified) maximum likelihood estimator $\hat{\theta}^{MLE}$ that a researcher would obtain if they knew the selection mechanism. We compute the Hessian and the outer product forms for the covariance matrix, and evaluate them at $\hat{\theta}^{MLE}$ to carry out the Information Matrix test. We report the rejection probability of this infeasible test in Table II, labeling it "IM rej." As can be seen from the table, for levels of $\gamma \in \{-.1, -.2\}$ the rejection probability is low, reaching at most 7.6% under Design 1; yet, Andrews and Shi (2013)'s test already shows non-trivial size distortions (16% and 37%, respectively). For $\gamma = -0.3, -.4$, the power is higher, reaching 91.6% under Design 1. For such settings, the size of Andrews and Shi (2013)'s test is substantially distorted (75% and 96%, respectively, against a 5% nominal level). In contrast, our test has correct size throughout all simulations, and maintains a power curve that is very similar to the one it displays in the case of correct model specification.

TABLE II

REJECTION PROBABILITIES OF SCORE AND MOMENT INEQUALITY TESTS

Design	Specification info.	Tests	Size	Power (values of $h/\sqrt{n}$ below)								
				0.013	0.025	0.038	0.050	0.063	0.076	0.088	0.101	0.113
Panel A: Correctly specified $(\gamma = 0)$												
Design 1:	Two Covariates											
		Score Test	0.028	0.058	0.154	0.432	0.732	0.916	0.996	1.000	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 4)$	0.070	0.106	0.222	0.430	0.722	0.896	0.976	0.998	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 3)$	0.066	0.102	0.214	0.410	0.714	0.892	0.974	0.998	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 2)$	0.058	0.088	0.202	0.394	0.690	0.886	0.974	0.998	1.000	1.000
Design 2:	Three Covariates											
		Score Test	0.018	0.050	0.150	0.390	0.708	0.904	0.984	1.000	1.000	1.000
Panel B: Misspecified $(\gamma = -0.1)$												
Design 1:	Two Covariates											
	$I(p_0||q_\theta^*) = $ 1.99e-04	Score Test	0.046	0.086	0.192	0.476	0.746	0.932	0.992	1.000	1.000	1.000
	IM rej. = 0.002	Moment Ineq. Test $(r_{1n} = 4)$	0.160	0.196	0.344	0.600	0.818	0.940	0.988	0.998	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 3)$	0.158	0.192	0.342	0.590	0.816	0.936	0.988	0.998	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 2)$	0.148	0.178	0.310	0.550	0.808	0.936	0.984	0.996	1.000	1.000
Design 2:	Three Covariates											
	$I(p_0||q_\theta^*) = $ 2.00e-04	Score Test	0.050	0.092	0.242	0.494	0.762	0.938	0.986	1.000	1.000	1.000
	IM rej.= 0.002											
Panel C: Misspecified $(\gamma = -0.2)$												
Design 1:	Two Covariates											
	$I(p_0||q_\theta^*) =$8.15e-04	Score Test	0.042	0.070	0.152	0.414	0.710	0.910	0.986	1.000	1.000	1.000
	IM rej. = 0.076	Moment Ineq. Test $(r_{1n} = 4)$	0.386	0.462	0.644	0.824	0.936	0.986	0.996	1.000	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 3)$	0.374	0.452	0.636	0.816	0.932	0.982	0.998	1.000	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 2)$	0.358	0.440	0.624	0.792	0.922	0.980	0.996	1.000	1.000	1.000
Design 2:	Three Covariates											
	$I(p_0||q_\theta^*) = $ 8.15e-04	Score Test	0.036	0.076	0.194	0.412	0.716	0.910	0.982	1.000	1.000	1.000
	IM rej.= 0.038											
Panel D: Misspecified $(\gamma = -0.3)$												
Design 1:	Two Covariates											
	$I(p_0||q_\theta^*) = $ 1.81e-03	Score Test	0.028	0.054	0.136	0.358	0.660	0.864	0.970	0.996	1.000	1.000
	IM rej. = 0.498	Moment Ineq. Test $(r_{1n} = 4)$	0.762	0.830	0.896	0.968	0.986	1.000	1.000	1.000	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 3)$	0.746	0.812	0.890	0.966	0.984	1.000	1.000	1.000	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 2)$	0.724	0.794	0.878	0.958	0.982	0.998	1.000	1.000	1.000	1.000
Design 2:	Three Covariates											
	$I(p_0||q_\theta^*) = $ 1.80e-03	Score Test	0.020	0.046	0.150	0.346	0.644	0.876	0.962	0.998	1.000	1.000
	IM rej. = 0.274											
Panel E: Misspecified $(\gamma = -0.4)$												
Design 1:	Two Covariates											
	$I(p_0||q_\theta^*) = $ 3.14e-03	Score Test	0.024	0.046	0.130	0.312	0.616	0.852	0.968	0.992	1.000	1.000
	IM rej. = 0.916	Moment Ineq. Test $(r_{1n} = 4)$	0.964	0.980	0.992	0.998	1.000	1.000	1.000	1.000	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 3)$	0.964	0.980	0.992	0.996	1.000	1.000	1.000	1.000	1.000	1.000
		Moment Ineq. Test $(r_{1n} = 2)$	0.954	0.974	0.988	0.996	1.000	1.000	1.000	1.000	1.000	1.000
Design 2:	Three Covariates											
	$I(p_0||q_\theta^*) = $ 3.14e-03	Score Test	0.030	0.054	0.148	0.372	0.658	0.862	0.972	0.996	1.000	1.000
	IM rej. = 0.760											

Note: The simulation results are based on samples of size $n = 7{,}017$ and 500 Monte Carlo repetitions. Panel A reports results for correctly specified models. Panels B-F report results for misspecified models with different values of γ. The second column reports the number of continuous covariates and the degree of misspecification measured by the KL divergence, and the rejection probability of an infeasible Information Matrix test.

TABLE III

COMPUTATIONAL TIME COMPARISONS FOR DESIGN 1 (IN SECONDS)

	Score Test	Moment Inequality Test		
		$r_{1n}=2$	$r_{1n}=3$	$r_{1n}=4$
Calculating the Statistic	0.25	8.75	30.04	73.62
Calculating the Critical Value	2.04e-4	48.4	247.31	818.23

Table III reports average computational time in seconds to calculate test statistics and critical values in Design 1 for 500 Monte Carlo replications on Boston University's computing cluster (with Intel Xeon Gold 6132 Processors and 192GB RAM). Our test statistic is 35-294 times faster to compute than Andrews and Shi's (2013), but the most substantial gain comes from calculation of the critical value: 0.0002 seconds for us, against 48-818 for Andrews and Shi (2013).

7. CONCLUSIONS

This paper is concerned with statistical inference in incomplete models with set valued predictions. Such models are typically partially identified, and can be misspecified. Misspecification can make the identification region of the model's parameters spuriously tight or even empty, raising a challenge for interpreting identification results, and can cause existing testing procedure to severely overreject. We propose to resolve these problems through an information-based method. Our method delivers a non-empty pseudo true set which can be interpreted as the set of minimizers of the researcher's ignorance about the true structure, as in White (1982). For any given parameter value, our inference method solves a convex program to find the density function that is closest to the data generating process with respect to the Kullback-Leibler information criterion. It then obtains the score of the likelihood function associated with this density and a Rao score test statistic. We show that the test statistic has an asymptotically pivotal distribution, is easy to compute, and does not require moment selection. The associated test has uniformly valid asymptotic size, is applicable to both correctly specified and misspecified models, and allows for discrete and continuous covariates. Monte Carlo simulations confirm the good computational and statistical properties of our proposed inference method.

APPENDIX A: PROOFS OF MAIN THEOREMS

Proof of Theorem 3.1. Part (i). As shown in (3.23), $L(\theta|x)$ is the optimal value function of a convex program. Below, we fix x and drop conditioning from L, p_0, q, and ν_θ to ease notation. Lemma B.2 in the Online Appendix delivers two key results. First, there is a collection of events $\mathcal{A}^{(*e)} \subseteq 2^{\mathcal{Y}}$ that does not depend on $\theta \in \Theta$, such that

$$\mathrm{core}(\nu_\theta(\cdot|x)) = \left\{ Q \in \mathcal{M}(\Sigma_Y, \mathcal{X}) : Q(A|x) \geq \nu_\theta(A|x), A \in \mathcal{A}^{(*e)} \right\}, \qquad \text{(A.1)}$$

where $\mathrm{core}(\nu_\theta(\cdot|x))$ is defined in (3.2), and the cardinality of the collection $\mathcal{A}^{(*e)}$ is the smallest among any collection of test sets guaranteeing (A.1). Hence, it suffices to verify the dominance condition in (3.2) for all $A \in \mathcal{A}^{(*e)}$ rather than for all $A \in \mathcal{C}$.[23]

Second, the problem

$$\begin{aligned} L(\theta) = \max_{q \in \Delta} \sum_{y \in \mathcal{Y}} p_0(y) \ln q(y) & \\ \text{s.t.} \quad \nu_\theta(A) \leq \sum_{y \in A} q(y),\ A \in \mathcal{A}^{(*e)}, & \end{aligned}$$

has a unique solution q^* with unique Lagrange multiplier vector λ^* associated with the constraints. We let $J = |\mathcal{A}^{(*e)}|$ and we denote sets in $\mathcal{A}^{(*e)}$ by $A_j, j = 1, \ldots, J$

Consider $V(t) = L(\theta + th)$ for $h \in \mathbb{R}^{d_\theta}$ and $t \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$. Note that q may be viewed as a vector because $\mathcal{Y}$ is finite. Below, we view $V(t)$ as the optimal value function of the convex program with objective function $f(q,t) = \sum_{y \in \mathcal{Y}} p_0(y) \ln q(y)$ and convex (affine) constraints $g_j(q,t) = \nu_{\theta+th}(A_j) - \sum_{y \in A_j} q(y)$, $j = 1, \ldots, J$. Note that Δ is compact and convex. Both f and g_j's are continuous and concave in q. Therefore, for any sequence $\{t_n\}$ with $t_n \downarrow 0$, the maximizer of $L(\theta + t_n h)$ exists. Furthermore, since the domain of the control variable and parameter $\Delta \times (-\epsilon, \epsilon)$ is bounded, the inf-boundedness assumption of Rockafellar (1984) holds. This ensures that the parametric optimization prob-

[23] Galichon and Henry (2011) call collections of sets with this property *core determining*. Applying results in Luo et al. (2025), Lemma B.2 in the Online Appendix shows that $\mathcal{A}^{(*e)}$ is an *exact core determining class*, i.e., it has smallest cardinality among core determining classes.

lem indexed by t is directionally stable in the sense of Gauvin and Janin (1990). Furthermore, their derivatives with respect to t are $f_t(q,t)=0$ and $g_{j,t}(q,t)=\nabla_\theta\nu_{\theta+th}(A_j)^\top h$, and they are continuous in (q,t) by assumption. Let $\mathcal{L}(q,\lambda,t)=f(q,t)+\sum_j^J\lambda_j g_j(q,t)$ be the Lagrangian. By Gauvin and Janin (1990, Corollary 4.2) and (q^*,λ^*) being unique, V is differentiable at $t=0$ and its derivative is given by

$$V'(0)=\tfrac{d}{dt}\mathcal{L}(q^*,\lambda^*,t)|_{t=0}=\sum_j^J\lambda_j^*\nabla_\theta\nu_\theta(A_j)^\top h. \tag{A.2}$$

Since this holds for any $h\in\mathbb{R}^{d_\theta}$, $L(\theta)$ is differentiable with

$$\nabla_\theta L(\theta)=\sum_j\lambda_j^*\nabla_\theta\nu_\theta(A_j). \tag{A.3}$$

Part (ii)-Eq.(3.21). In what follows we construct a score function. Let $M=|\mathcal{Y}|$, and order the elements of $\mathcal{Y}$ as $y_1,y_2,\dots,y_M$. Let $\mathcal{J}=\{j\in\{1,\dots,J\}:\sum_{\tilde{y}\in A_j}q^*(\tilde{y})=\nu_\theta(A_j)\}$ be the set of active constraints, and let $\mathcal{J}^c=\{1,\dots,J\}\setminus\mathcal{J}$ collect slack constraints. For each $y\in\mathcal{Y}$, let $\mathcal{J}(y)=\{j\in\mathcal{J}:y\in A_j\}$ collect the indices associated with the active constraints such that y belongs to A_j. By the Karush-Kuhn-Tucker conditions, differentiating $\mathcal{L}$ with respect to q and evaluating it at q^* yields

$$\tfrac{p_0(y_m)}{q^*(y_m)}+\sum_{j\in\mathcal{J}(y_m)}\lambda_j^*=0,\quad m=1,\dots,M. \tag{A.4}$$

For each $y\in\mathcal{Y}$, let $e_{\mathcal{J}(y)}\in\{0,1\}^J$ be a vector whose j-th component is 1 if $j\in\mathcal{J}(y)$ and 0 otherwise. Then, the system of equations (A.4) can be written as

$$B\lambda^*=r, \tag{A.5}$$

where B is an M-by-J matrix and r is an M-by-1 vector with m-th row defined as follows

$$[B]_{[m,\cdot]}=-e'_{\mathcal{J}(y_m)},\qquad [r]_m=\tfrac{p_0(y_m)}{q^*(y_m)}. \tag{A.6}$$

By the complementary slackness conditions, $\lambda_j^* = 0$ for any $j \in \mathcal{J}^c$. Hence, (A.5) can be reduced to a system of M equations with $S = |\mathcal{J}|$ unknowns. Eliminate the columns of B corresponding to $j \in \mathcal{J}^c$ and let $\tilde{B}$ denote the resulting submatrix of B. Similarly, eliminate the components of λ corresponding to $j \in \mathcal{J}^c$ and let $\tilde{\lambda}^*$ denote the resulting subvector. (A.5) can be rewritten as $\tilde{B}\tilde{\lambda}^* = r$, where $\tilde{B}$ is a $M \times S$ matrix whose columns are the representers $\{b_{A_j}, j \in \mathcal{J}\}$ of the active constraints. By Lemma B.2 (iv), the vectors $\{b_{A_j}, j \in \mathcal{J}\}$ are linearly independent. Hence, $\tilde{\lambda}^*$ solves equation $\tilde{B}\tilde{\lambda}^* = r$ uniquely, and there exists an S-by-M matrix C such that $\tilde{\lambda}^* = Cr$. Let E_θ be a $d \times S$ matrix that stacks the column vectors $\{\nabla_\theta \nu_\theta(A_j), j \in \mathcal{J}\}$. Then,

$$\nabla_\theta L(\theta) = \sum_j^J \lambda_j^* \nabla_\theta \nu_\theta(A_j) = \sum_{j \in \mathcal{J}} \lambda_j^* \nabla_\theta \nu_\theta(A_j) = E_\theta \tilde{\lambda}^* = E_\theta C r$$

by (A.3) and $\tilde{\lambda}^* = Cr$. Recall that r is defined in (A.6). Hence, $\nabla_\theta L(\theta)$ can be written as

$$\nabla_\theta L(\theta) = \sum_{m=1}^M p_0(y_m) \frac{[E_\theta C]_{[\cdot,m]}}{q^*(y_m)}, \tag{A.7}$$

where $[\cdot]_{[\cdot,m]}$ selects the m-th column of its argument. Now let $s_\theta(y_m) = \frac{[E_\theta C]_m}{q^*(y_m)}$ and recall that so far we dropped conditioning on x and dependence on $p_{0,y|x}$. (A.7) therefore shows that $\nabla_\theta L(\theta|x) = \mathbb{E}[s_\theta(Y|X; p_{0,y|x}))|X = x]$, and s_θ's square integrability follows from $q^*(y) > 0$, $\nabla_\theta \nu_\theta(A_j|X)$ being square integrable for all j, and $\mathcal{Y}$ being a finite set.

Part (ii)-(3.22). By law of iterated expectations and dominated convergence theorem,

$$\begin{aligned} \mathbb{E}[s_\theta(Y|X; p_{0,y|x})] &= \mathbb{E}[\mathbb{E}[s_\theta(Y|X; p_{0,y|x})|X]] \\ &= \mathbb{E}\left[\tfrac{\partial}{\partial\theta} L(\theta|X)\right] = \tfrac{\partial}{\partial\theta}\mathbb{E}[L(\theta|X)] = \tfrac{\partial}{\partial\theta} L(\theta) = 0, \end{aligned}$$

for any $\theta \in \Theta^*(p_0)$, where the last equality follows from the first-order condition for maximizing $\theta \mapsto L(\theta)$ and the fact that $\Theta^*(p_0) \subset \operatorname{int}\Theta$. *Q.E.D.*

Proof of Theorem 3.2. The result follows from Lemma B.3, which establishes joint continuity of $\phi(\cdot,\cdot)$ in (θ, p), and the use of Berge's maximum theorem, observing that $\mathfrak{q}_\theta$ does not depend on p. *Q.E.D.*

We next turn to the proof of Theorem 3.3. To establish the result, we first show that the expected score satisfies an asymptotic orthogonality condition with respect to the nuisance parameter. This result ensures that the score statistic's limiting distribution is insensitive to the nonparametric estimation of the conditional choice probability. Below, let $q_{\theta,h_x,y|x} \in \mathfrak{q}_{\theta,x}$ be indexed by the structural parameter θ and equilibrium selection $h_x = \{h_{x,u},\ u \in \mathcal{U}\}$, with $h_{x,u} \in \mathcal{S}(x,u;\theta)$ and $\mathcal{S}(x,u;\theta)$ the set of all conditional densities of $Y|X,U$ such that, for any (x,u), its conditional support is $G(u|x;\theta)$. Let $\mathcal{S}(\theta) = \{\mathcal{S}(x,u;\theta), x \in \mathcal{X}, u \in \mathcal{U}\}$.

LEMMA A.1: *Suppose Assumptions 1 and 2 hold. Then,*

$$D(\theta^*, p_{0,y|x})[p_{y|x} - p_{0,y|x}] = 0. \tag{A.8}$$

Proof of Lemma A.1: We rely on an application of Newey (1994, Proposition 2). Let

$$\rho(x,\theta,h_x) = \mathbb{E}_{P_0}[\ln q_{\theta,h_x,y|x}(Y|X)|X = x]. \tag{A.9}$$

Let $h_x(p) = [h_{x,u}(p), u \in \mathcal{U}]$, $h_{x,u}(p) \in \mathcal{S}(x,u;\theta)$, be the selection such that $q_{\theta,h_x(p),y|x} = q^*_{\theta,y|x}$ when p replaces p_0 in (3.9); this selection exists by Artstein (1983). Solving the KL projection problem in (3.8) (with p replacing p_0) is equivalent to maximizing out the equilibrium selection. Therefore, the function valued parameter $h(p) \in \mathcal{S}(\theta)$ solves $h(p) = \arg\max_{\tilde{h}\in\mathcal{S}(\theta)} \mathbb{E}_{P_0}[\rho(x,\theta,\tilde{h})]$, where the expectation is taken with respect to the true DGP distribution P_0 and the dependence of $h(\cdot)$ on p results from the KL projection step. Arguing as in Newey (1994), for a path P_τ, denoting $h(\tau) = h(p_\tau)$, we have that $\mathbb{E}_{P_0}[\rho(x,\theta,h(\tau))] \le \max_{\tilde{h}\in\mathcal{S}(\theta)} \mathbb{E}_{P_0}[\rho(x,\theta,\tilde{h})]$ and hence $\mathbb{E}_{P_0}[\rho(x,\theta,h(\tau))]$ is maximized at $\tau = 0$. The first order conditions for this maximum are $\partial\mathbb{E}[\rho(x,\theta,h(\tau))]/\partial\tau = 0$ for all θ.

Differentiating one more time with respect to θ and using the law of iterated expectations,

$$0 = \frac{\partial^2}{\partial\tau\partial\theta}\mathbb{E}_{P_0}[\rho(x,\theta^*,h_x(\tau))]\big|_{\tau=0} = \frac{\partial}{\partial\tau}\mathbb{E}_{P_0}\left[\frac{\partial}{\partial\theta}\mathbb{E}_{P_0}[\ln q^*_{\theta,y|x}(Y|X)|X=x]\right]\Bigg|_{\tau=0}$$
$$= \frac{\partial}{\partial\tau}\mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}[s_\theta(Y|X;p_{\tau,y|x})|X=x]\right]\big|_{\tau=0} = \frac{\partial}{\partial\tau}\mathbb{E}_{P_0}[s_{\theta^*}(Y|X;p_{\tau,y|x})]\big|_{\tau=0},$$

where (3.21) yields the third equality. Hence, the pathwise derivative is zero. *Q.E.D.*

Proof of Theorem 3.3. Let us write the left-hand side of (3.31) as

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x}) = \mathbb{G}_{n,\theta^*}(\hat{p}_{n,y|x}) + \sqrt{n}\mathbb{E}[s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x})]$$
$$= \mathbb{G}_{n,\theta^*}(p_{0,y|x}) + (\mathbb{G}_{n,\theta^*}(\hat{p}_{n,y|x}) - \mathbb{G}_{n,\theta^*}(p_{0,y|x})) + \sqrt{n}\mathbb{E}[s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x})]. \quad \text{(A.10)}$$

By Assumption 3-(iii), $\mathbb{G}_{n,\theta^*}(p_{0,y|x}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n s_{\theta^*}(Y_i|X_i;p_{0,y|x})) \xrightarrow{d} N(0,\Sigma_{\theta^*})$. Furthermore, by Assumptions 3 (ii) and 3 (iv), $\mathbb{G}_{n,\theta^*}(\hat{p}_{n,y|x}) - \mathbb{G}_{n,\theta^*}(p_{0,y|x}) = o_p(1)$.

Let $r_n \equiv \mathbb{E}[m_{\theta^*}(X_i;\hat{p}_{n,y|x})] - \mathbb{E}[m_{\theta^*}(X_i;p_{0,y|x})] - D(\theta^*,p_{0,y|x})[\hat{p}_{n,y|x} - p_{0,y|x}]$. Then,

$$\begin{aligned}\mathbb{E}[s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x})] &= \mathbb{E}[\mathbb{E}[s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x})|X=x]]\\ &= \mathbb{E}[m_{\theta^*}(X_i;p_{0,y|x})] + \mathbb{E}[m_{\theta^*}(X_i;\hat{p}_{n,y|x}) - m_{\theta^*}(X_i;p_{0,y|x})]\\ &= \mathbb{E}[s_{\theta^*}(Y_i|X_i;p_{0,y|x})] + D(\theta^*,p_{0,y|x})[\hat{p}_{n,y|x} - p_{0,y|x}] + r_n,\end{aligned}$$

where the first equality follows from the law of iterated expectations, the second equality follows from the definition of m_θ, and the third equality follows from the law of iterated expectations and the definition of r_n. As $\theta^* \in \Theta^*(p_0)$, $\mathbb{E}[s_{\theta^*}(Y_i|X_i;p_{0,y|x})] = 0$. By Lemma A.1, $\sqrt{n}D(\theta^*,p_{0,y|x})[\hat{p}_{n,y|x} - p_{0,y|x}] = 0$. Finally, using Assumptions 2 and 3 (ii),

$$|\sqrt{n}r_n| \le \sqrt{n}\|r_n\|_{L^2_P} \le \sqrt{n}c\|\hat{p}_{n,y|x} - p_{0,y|x}\|^2_{\mathcal{H}} = o_p(1).$$

Hence, by the triangle inequality,

$$|\sqrt{n}\mathbb{E}[s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x})]| \;\leq\; |\sqrt{n}\mathbb{E}[s_{\theta^*}(Y_i|X_i;p_{0,y|x})]| \;+\; |\sqrt{n}r_n| \;=\; o_p(1). \quad \text{(A.11)}$$

Combining (A.10)-(A.11) yields the result in (3.31). *Q.E.D.*

Proof of Corollary 3.1. For the case that Σ_{θ^*} is nonsingular, standard arguments, Assumption 3, and (3.35) yield $T_n(\theta^*) \xrightarrow{d} \mathsf{J}$, with $\mathsf{J} \sim \chi^2_{d_\theta}$. For the case that Σ_{θ^*} is singular, let $\zeta \in \mathbb{R}^{d_\theta}$ be a random vector such that $\zeta = \eta + \nu$, with $\eta \perp\!\!\!\perp \nu$, $\eta \sim N(0, \Sigma_{\theta^*})$, and $\nu \sim N(0, \varepsilon\Psi_{\theta^*})$, where Ψ_{θ^*} is the population analog of $\hat{\Psi}_{n,\theta^*}$ (see (3.34) and subsequent explanation of notation). Let $\tilde{\Sigma}_{\theta^*} = \Sigma_{\theta^*} + \varepsilon\Psi_{\theta^*}$. It follows from standard arguments that $T_n \xrightarrow{d} \eta^\top \tilde{\Sigma}^{-1}_{\theta^*}\eta$, and that $\zeta^\top \tilde{\Sigma}^{-1}_{\theta^*}\zeta \sim \mathsf{J}$. Next, let $K = \{x \in \mathbb{R}^{d_\theta} : x^\top \tilde{\Sigma}^{-1}_{\theta^*} x \leq c_{d_\theta,\alpha}\}$, and note that this set is convex and symmetric. By Anderson's Lemma (van der Vaart and Wellner, 1996, Lemma 3.11.4),

$$1 - \alpha = P(\zeta^\top \tilde{\Sigma}^{-1}_{\theta^*}\zeta \leq c_{d_\theta,\alpha}) = P(\eta + \nu \in K) \leq P(\eta \in K) = P(\eta^\top \tilde{\Sigma}^{-1}_{\theta^*}\eta \leq c_{d_\theta,\alpha}).$$

Then $\limsup_{n\to\infty} P(T_n(\theta^*) > c_{d_\theta,\alpha}) = P(\eta^\top \tilde{\Sigma}^{-1}_{\theta^*}\eta > c_{d_\theta,\alpha}) \leq P(\zeta^\top \tilde{\Sigma}^{-1}_{\theta^*}\zeta > c_{d_\theta,\alpha}) = \alpha$. *Q.E.D.*

Proof of Theorem 3.4. Let $\{p_{0n}, \theta_n^*\} \in \{(p, \vartheta^*) : p \text{ is the Radon-Nykodim derivative of } P \in \mathcal{P}, \vartheta^* \in \Theta^*(p)\}$ be a sequence such that:

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\vartheta^*\in\Theta^*(p)} P(\vartheta^* \in CS_n) = \liminf_{n\to\infty} P_n(\theta_n^* \in CS_n),$$

with CS_n defined in (3.36). Let $\{l_n\}$ be a subsequence of $\{n\}$ such that

$$\liminf_{n\to\infty} P_n(\theta_n^* \in CS_n) = \lim_{n\to\infty} P_{l_n}(\theta^*_{l_n} \in CS_{l_n})).$$

Then there is a further subsequence $\{a_n\}$ of $\{l_n\}$ such that

$$\lim_{a_n\to\infty} \Sigma_{\theta^*_{a_n}} = \Sigma^* \in \mathbb{S},$$

where $\mathbb{S}$ is the collection of positive semi-definite $d_\theta \times d_\theta$ matrices. To avoid multiple subscripts, with some abuse of notation we write (P_n, θ_n^*) to refer to $(P_{a_n}, \theta_{a_n}^*)$. We establish the claim by showing that along the subsequence (P_n, θ_n^*), the results in Theorem 3.1, Lemma A.1, and Theorem 3.3 continue to hold.

For Theorem 3.1, note first that the collection of events $\mathcal{A}^{(*e)}$ does not depend on (P_n, θ_n^*), as can be seen in the proof of Lemma B.2-(i). Second, parts (ii) and (iii) of Lemma B.2 continue to hold along the subsequence (P_n, θ_n^*) under the uniform version of Assumption 1 stated in Theorem 3.4.

For Lemma A.1, we again note that the result holds uniformly over $\mathcal{P}$, under the uniform version of Assumptions 1, 2, and 3 (ii) stated in Theorem 3.4.

For Theorem 3.3, under the uniform version of Assumptions 1, 2, and 3 stated in Theorem 3.4, we have that by Assumption 3, $\mathbb{G}_{n,\theta_n^*}(p_{0n,y|x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n s_{\theta_n^*}(Y_i|X_i; p_{0n,y|x})) \xrightarrow{d} N(0, \Sigma^*)$, and $\mathbb{G}_{n,\theta_n^*}(\hat{p}_{n,y|x}) - \mathbb{G}_{n,\theta_n^*}(p_{0n,y|x}) = o_{P_n}(1)$. Arguing as in the proof of Theorem 3.3, (A.10)-(A.11) continue to hold along the sequence (P_n, θ_n^*), and therefore

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n s_{\theta_n^*}(Y_i|X_i; \hat{p}_{n,y|x}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n s_{\theta_n^*}(Y_i|X_i; p_{0,y|x}) + o_{\mathcal{P}}(1) \xrightarrow{d} N(0, \Sigma^*).$$

The final result follows arguing as in the proof of Corollary 3.1. *Q.E.D.*

APPENDIX: REFERENCES

AKAIKE, HIROTOGU (1973): *Information Theory and an Extension of the Maximum Likelihood Principle*, Budapest, Hungary: Akadé Kiadó, 267–281. [3]

ANDREWS, DONALD W. K. AND PANLE JIA BARWICK (2012): “Inference for parameters defined by moment inequalities: a recommended moment selection procedure,” *Econometrica*, 80, 2805–2826. [20]

ANDREWS, DONALD W. K. AND SOONWOO KWON (2024): “Misspecified Moment Inequality Models: Inference and Diagnostics,” *Review of Economic Studies*, 91, 45–76. [4, 21]

ANDREWS, DONALD W. K. AND XIAOXIA SHI (2013): “Inference based on conditional moment inequalities,” *Econometrica*, 81, 609–666. [5, 15, 21, 31, 32, 33, 35]

ARTSTEIN, Z. (1983): “Distributions of random sets and random selections,” *Israel Journal of Mathematics*, 46, 313–324. [2, 8, 39]

AUBIN, J.-P. AND H. FRANKOWSKA (2009): *Set-Valued Analysis*, Birkhäuser Boston, Inc., Boston, MA. [17]

BARSEGHYAN, LEVON, MAURA COUGHLIN, FRANCESCA MOLINARI, AND JOSHUA C. TEITELBAUM (2021): "Heterogeneous Choice Sets and Preferences," *Econometrica*, 89, 2015–2048. [11, 13]

BERESTEANU, ARIE, ILYA MOLCHANOV, AND FRANCESCA MOLINARI (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79, 1785–1821. [2, 7, 9]

BERRY, STEVEN T. (1992): "Estimation of a Model of Entry in the Airline Industry," *Econometrica*, 60, 889–917. [7]

BERRY, STEVEN T. AND GIOVANNI COMPIANI (2022): "An Instrumental Variable Approach to Dynamic Models," *Review of Economic Studies*, 90, 1724–1758. [13]

BONTEMPS, CHRISTIAN AND ROHIT KUMAR (2020): "A geometric approach to inference in set-identified entry games," *Journal of Econometrics*, 218, 373–389. [23]

BUGNI, FEDERICO A., IVAN A. CANAY, AND XIAOXIA SHI (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8, 1–38. [5, 15, 21]

CHEN, SHUOWEN AND HIROAKI KAIDO (2023): "Robust Tests of Model Incompleteness in the Presence of Nuisance Parameters," available at https://arxiv.org/abs/2208.11281. [14]

CHEN, XIAOHONG, TIMOTHY M. CHRISTENSEN, AND ELIE TAMER (2018): "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86, 1965–2018. [3, 4, 11, 21, 28, 29]

CHEN, XIAOHONG, LARS PETER HANSEN, AND PETER G. HANSEN (2024): "Robust inference for moment condition models without rational expectations," *Journal of Econometrics*, 243, 105653. [4]

CHEN, XIAOHONG, OLIVER LINTON, AND INGRID VAN KEILEGOM (2003): "Estimation of Semiparametric Models When the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608. [19]

CHEN, XIAOHONG, ELIE TAMER, AND ALEXANDER TORGOVITSKY (2011): "Sensitivity Analysis in Semiparametric Likelihood Models," Cowles Foundation Discussion Paper No.1836. [3, 4, 11]

CILIBERTO, FEDERICO AND ELIE TAMER (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77, 1791–1828. [8]

EPSTEIN, L. G., H. KAIDO, AND K. SEO (2016): "Robust confidence regions for incomplete models," *Econometrica*, 84, 1799–1838. [19]

FORNERON, JEAN-JACQUES (2023): "Noisy, Non-Smooth, Non-Convex Estimation of Moment Condition Models," available at https://arxiv.org/pdf/2301.07196.pdf. [24]

GALICHON, ALFRED AND MARC HENRY (2011): "Set Identification in Models with Multiple Equilibria," *The Review of Economic Studies*, 78, 1264–1298. [36]

GAUVIN, JACQUES AND ROBERT JANIN (1990): "Directional derivative of the value function in parametric optimization," *Annals of Operations Research*, 27, 237–252. [14, 37]

GU, JIAYING, THOMAS RUSSELL, AND THOMAS STRINGHAM (2025): "Counterfactual Identification and Latent Space Enumeration in Discrete Outcome Models," *The Review of Economic Studies*. [13]

HANSEN, BRUCE E. AND SEOJEONG LEE (2021): "Inference for iterated GMM under misspecification," *Econometrica*, 89, 1419–1447. [2]

HONORÉ, BO E. AND ELIE TAMER (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629. [6]

KAIDO, HIROAKI, FRANCESCA MOLINARI, AND JÖRG STOYE (2019): "Confidence Intervals for Projections of Partially Identified Parameters," *Econometrica*, 87, 1397–1432. [5, 15, 21]

KLINE, BRENDAN AND ELIE TAMER (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7, 329–366. [5, 26, 27, 30, 31]

KOOPMANS, T. C. AND O. REIERSOL (1950): "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics*, 21, 165–181. [2]

LI, LIXIONG, DÉSIRÉ KÉDAGNI, AND ISMAËL MOURIFIÉ (2024): "Discordant relaxations of misspecified models," *Quantitative Economics*, 15, 331–379. [2, 5, 15, 23]

LUO, YE, KIRILL PONOMAREV, AND HAI WANG (2025): "Selecting Inequalities for Sharp Identification in Models with Set-Valued Predictions," . [13, 14, 23, 36]

MOLCHANOV, ILYA AND FRANCESCA MOLINARI (2018): *Random Sets in Econometrics*, Econometric Society Monograph Series, Cambridge University Press, Cambridge UK. [8]

MOLINARI, FRANCESCA (2020): "Microeconometrics with Partial Identification," in *Handbook of Econometrics*, Volume 7A, ed. by Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, and Rosa L. Matzkin, Amsterdam: Elsevier, 355–486. [2, 6]

NEWEY, WHITNEY K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382. [3, 18, 39]

ROCKAFELLAR, R. T. (1984): *Directional differentiability of the optimal value function in a nonlinear programming problem*, Berlin, Heidelberg: Springer Berlin Heidelberg, 213–226. [36]

ROCKAFELLAR, R. TYRRELL AND ROGER J.-B. WETS (2005): *Variational Analysis, Second Edition*, Springer-Verlag, Berlin. [16, 17]

STOYE, JÖRG (2020): "A Simple, Short, but Never-Empty Confidence Interval for Partially Identified Parameters," available at https://arxiv.org/abs/2010.10484. [4]

TAMER, ELIE (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70, 147–165. [6]

VAN DER VAART, A. W. AND JON A. WELLNER (1996): "Weak Convergence and Empirical Processes, With Applications to Statistics," *Springer Series in Statistics*. [19, 41]

WHITE, HALBERT (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [2, 3, 9, 10, 33, 35]

——— (1996): *Estimation, inference and specification analysis*, 22, Cambridge university press. [9, 10]