ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Kaido, Hiroaki; Molinari, Francesca

Working Paper

Information based inference in models with set-valued predictions and misspecification

cemmap working paper, No. CWP02/24

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Kaido, Hiroaki; Molinari, Francesca (2024) : Information based inference in models with set-valued predictions and misspecification, cemmap working paper, No. CWP02/24, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.47004/wp.cem.2024.0224

This Version is available at:
https://hdl.handle.net/10419/284150

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

cemmap
centre for microdata methods and practice

Information based inference in models with set-valued predictions and misspecification

Hiroaki Kaido
Francesca Molinari

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP02/24

UKRI Economic and Social Research Council

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

Over the last twenty years, a rich literature has emerged to provide methods to carry out inference in partially identified models under the assumption of correct model specification. In some cases this assumption is plausible, as partial identification often results from reducing the number of suspect assumptions maintained in counterpart point identified models.

Hiroaki Kaido: hkaido@bu.edu

Francesca Molinari: fm72@cornell.edu

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

Yet, even in partially identified models, rarely is inference based on the empirical evidence alone: researchers routinely impose exogeneity assumptions, behavioral restrictions, distributional and functional form specifications, etc. These conditions are merely approximations of complex social and economic phenomena, and hence subject to misspecification error. When a partially identified model is misspecified, at least three problems may occur (see Molinari, 2020, and references therein): (i) the parameters' sharp identification region may be empty or spuriously tight;[1] (ii) confidence sets constructed assuming correct model specification may (severely) undercover; and (iii) their tightness may be misinterpreted as highly informative data. Each of these problems has a counterpart in point identified models that, for maximum likelihood, least squares, and GMM estimators, has been addressed in the econometrics literature since at least White (1982).[2] Yet, development of a toolkit for inference in misspecified partially identified models is just starting.

We contribute to this nascent literature by proposing a novel information theoretic inference method in the spirit of White (1982) that is robust to misspecification, easy to implement, valid both for correctly and incorrectly specified models, and able to exploit all information provided by variation in discrete and continuous regressors. 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), and hence yield 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 Molinari, 2020, for a review of these models).

The method that we propose is based on adapting the textbook one for (point identified) models that predict a singleton conditional distribution for $Y|X$, to models that predict a

[1] The *sharp identification region* is the set of parameters that are observationally equivalent, i.e., can generate the same distribution of observables as the one in the data, for some DGP consistent with the maintained assumptions.

[2] See, e.g., Gallant and White (1988), Hall and Inoue (2003), Hansen and Lee (2021), and references therein.

set of distributions for $Y|X$. Our first step is to characterize the set of model predicted distributions. To do so, we leverage a result in random set theory due to Artstein (1983) to ensure that we collect exclusively the distributions consistent with all maintained assumptions. This is especially important in the context of misspecification, as using only a subset of model implications may yield misleading conclusions (Kédagni et al., 2021, Molinari, 2020, Beresteanu et al., 2011). In the second step, we define a never-empty pseudo-true set for the parameter vector θ characterizing the model. This is the collection of minimizers of a Kullback-Leibler (KL) information criterion that measures the divergence of the set of model-predicted distributions from the distribution of the observed data. This pseudo-true set has a similar information theoretic interpretation as originally proposed by Akaike (1973) and White (1982). It shrinks to the pseudo-true parameter vector in White (1982) if the assumptions are augmented so that the model predicts a single distribution. 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. We show that this profiling step can be 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 the profiled likelihood 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 (the bandwidth/sieves order of this estimator, and a regularization constant for a covariance matrix estimator, are the only tuning parameters we use). We leverage classic results in the semiparametric inference literature, specifically Newey (1994), to establish that an orthogonality property holds. Provided the rate of convergence of the nonparametric 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 the pseudo-true set. We invert the test to construct a confidence set and show that it is *robust to misspecification*: it covers each element of the pseudo-true set with asymptotic probability at least equal to the nominal level $1-\alpha$, uniformly over a large class of DGPs, both when the observed covariates have a discrete distribution with finite support and when they have a continuous distribution. The confidence set incorporates all information the covariates provide. The steps required to implement our method are the same both when the model is correctly specified and when it is misspecified, and resemble familiar ones based on the score of the likelihood function in point identified models.

Related Literature. The vast literature on inference in partially identified models (recently surveyed in Canay and Shaikh, 2017, Molinari, 2020) has considered questions related to misspecification. Early works (Manski, 2003, Ponomareva and Tamer, 2011) show that identification regions may be empty if the model is misspecified. Kaido and White (2013) study parametrically-misspecified moment inequality models, and provide a consistent estimator for a pseudo-true set and its rate of convergence. Several papers provide tests for correct specification; e.g., Guggenberger et al. (2008) and Bontemps et al. (2012) for linear moment (in)equality models, and Romano and Shaikh (2008), Andrews and Guggenberger (2009), Galichon and Henry (2009), and Andrews and Soares (2010) for general moment inequality models. Bugni et al. (2015) propose more powerful model specification tests that resemble the J-test for point identified GMM models. Bugni et al. (2012) show that with local misspecification, confidence sets constructed under the assumption of correct specification fail to asymptotically satisfy their nominal coverage requirement.

Only a few papers have put forth tools for construction of confidence sets that are valid in the presence of misspecification, in the sense of 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 (2022) show that model misspecification can lead to spuriously tight confidence sets while statistical tests (e.g., one proposed by Bugni et al. (2015) and one that they propose) have low power at detecting misspecification. They propose a notion of pseudo true set, a specification test, and an inference method for partially identified models defined by a *finite number of unconditional* moment inequalities. They obtain a

confidence set by test inversion that is uniformly valid in the presence of misspecification, using a test statistic that 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) focuses on the narrower class of interval identified scalar parameters, with finite number of moment conditions yielding the upper/lower bound and asymptotic normality for the estimators of these bounds. He obtains a valid and never-empty confidence interval that is free of tuning parameters and very simple to compute.

In contrast to these papers, our method is applicable to models for conditional density functions of outcome variables given discrete and continuous covariates. Allowing for continuously distributed covariates is important: they are commonplace in practice and may deliver substantial identifying information. Yet, existing empirical applications with partial identification rely on discretizing the covariates to obtain unconditional moment inequalities (e.g., Ciliberto and Tamer, 2009, Kline and Tamer, 2016, Dickstein and Morales, 2018). Doing so may distort the original model, and lead to loss of identifying information. Under the assumption of correct specification, one may use inference methods designed for conditional moment inequalities (e.g., Andrews and Shi, 2013). Such methods, however, rely on instrument functions to transform the conditional moments in an uncountable collection of unconditional ones, making their implementation challenging. We bypass this problem by evaluating directly the contribution of each observation to the score function. Our population pseudo-true set and our inference method are insensitive to which inequalities one uses to characterize the sharp collection of model implied distributions for $Y|X$. This is in contrast with much of the related literature, where moment selection is often required either for computational tractability or as part of the inference procedure, but may have substantial implications on the population region that the researcher targets (Kédagni et al., 2021) and on the properties of confidence sets (e.g., Andrews and Soares, 2010, Andrews and Shi, 2013, Bugni et al., 2017, Kaido et al., 2019).

Our method leverages results in random set theory (see Molchanov and Molinari, 2018, for a review), and contributes to likelihood-based inference approaches in partial identification. Among these approaches, Chen et al. (2011) consider correctly specified partially

identified semiparametric likelihood models and use sieve approximations for the infinite-dimensional parameters underlying the model. In contrast, we profile out the infinite-dimensional nuisance parameters through a (finite) convex program.[3] Chen et al. (2018) propose confidence sets for the identified set in correctly specified models, that are contour sets of a likelihood-based criterion function using cutoffs that are computed via Monte Carlo simulations from the quasi-posterior distribution of the criterion.[4] Chen et al. (2021) show that Chen et al. (2018)'s method delivers valid confidence sets for the pseudo true set in a class of dynamic models that are globally misspecified when subjective beliefs differ from the often maintained rational expectations assumption. Kaido and Zhang (2019) analyze some of the models that we consider, but under the assumption of correct model specification. They use the least favorable pairs (LFPs) of distributions studied in the robust statistics literature to build likelihood-ratio tests in incomplete models. Chen and Kaido (2023) introduce a score function derived from these LFPs, to test the assumption that the model predicts a unique distribution for $Y|X$. Their test is based on a Rao's score statistic that is distinct from ours both because under their null the model is point identified, and because they assume correct specification.

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

---

[3] Christensen and Connault (2023) provide tools for the related but distinct question of characterizing the sensitivity of counterfactuals to the distributional assumptions imposed on the latent variables of the model. They allow such distributions to span a nonparametric neighborhood of the parametric specification, and eliminate the infinite-dimensional nuisance parameter via a convex program of fixed dimension. In contrast, we focus on inference for the parameter vector characterizing the model (from which one can derive –possibly conservative– inference on counterfactuals), and allow for any type of misspecification.

[4] As the identified set in Chen et al. (2018) is given by the collection of maximizers of a properly defined likelihood function, it naturally yields a non-empty pseudo-true identified set under misspecification, which should be consistently estimated by the confidence sets that they propose.

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).[5] Assume that 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 let $p_0 \equiv \{p_{0,y|x}, x \in \mathcal{X}\}$.

Suppose the researcher posits: (i) restrictions on the joint behavior of (Y, X, U), such as, e.g., equilibrium or optimality conditions, which are expressed through functions known up to finite dimensional parameter vector; (ii) that the family of distributions for the latent variables U is known up to finite dimensional parameter vector. Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ denote the combined parameter vector, and (omitting specific notation for subvectors of θ) $\{F_\theta : \theta \in \Theta\}$ the family of distributions for U. Assume that F_θ is independent of X.[6]

We consider models with a structure that associates with each element of $\mathcal{U} \times \mathcal{X} \times \Theta$ a *set of predicted outcomes* through a closed valued and measurable correspondence $G : \mathcal{U} \times \mathcal{X} \times \Theta \mapsto \mathcal{Y}$.[7] 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. The next example clarifies notation and is used throughout the paper to illustrate results. More examples are provided in the Online Appendix[8] and in Molinari (2020, Sections 2 and 3).

Example 1 (Static entry game)**.** Consider a two player entry game as in Tamer (2003), where each player $i = 1, 2$ can choose to enter ($Y_i = 1$) or to stay out of the market ($Y_i = 0$). Let (X_1, X_2) and $(U_1, U_2) \sim F_\theta$ denote, respectively, observable and unobservable payoff shifters. Let player's payoffs be given by $\pi_j = Y_j(X_j\beta_j + \delta_j Y_{(3-j)} + U_j),\ j = 1, 2$, with $\delta_1 \leq 0, \delta_2 \leq 0$ the interaction effects and $(\beta_1, \beta_2, \delta_1, \delta_2)$ part of θ. Let each player enter

[5] 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 $(\mathcal{S}, \Sigma_\mathcal{S})$.

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

[7] 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 $\mathbb{R}^d$, $G^{-1}(K) = \{\omega \in \Omega : G(\omega) \cap K \neq \emptyset\} \in \mathfrak{F}$ (Molchanov and Molinari, 2018, Definition 1.1).

[8] These are: discrete choice models with unobserved heterogeneity in choice sets (Barseghyan et al., 2021) and panel dynamic discrete choice models (Heckman, 1978, Honoré and Tamer, 2006).

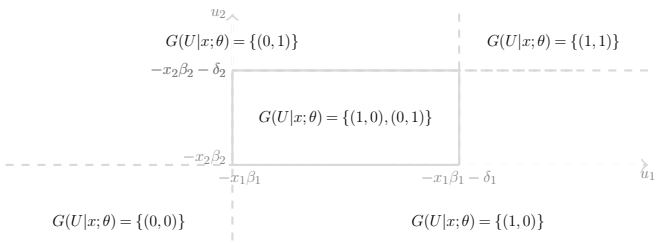


FIGURE 1.—Stylized depiction of $G(\cdot|x;\theta)$ in Example 1 with $\delta_1 < 0, \delta_2 < 0$.

the market if and only if $\pi_j \geq 0$. Given $\theta \in \Theta$ and $x \in \mathcal{X}$, the model has multiple pure strategy Nash equilibria (PSNE),[9] depicted in Figure 1 as a function of (u_1, u_2) (Tamer, 2003, Ciliberto and Tamer, 2009). In the notation of this paper, the set of PSNE is the measurable correspondence $G(\cdot|x;\theta)$ (Beresteanu et al., 2011, Proposition 3.1), with:

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

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

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

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

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

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 $g(U|x;\theta)$. □

[9] Multiple equilibria occur also, e.g., under different solution concepts (Aradillas-Lopez and Tamer, 2008, Magnolfi and Roncoroni, 2023) and in network formation models (de Paula et al., 2018, Sheng, 2020).

3. INFORMATION-BASED INFERENCE ROBUST TO MISSPECIFICATION

We characterize the set of model-implied probability density functions for $Y|X$ and use it to define a *model* and the notions of its *correct specification* and *misspecification*. We use the Kullback-Leibler (KL) divergence measure of this set of density functions from the (population) density function of the data to obtain a pseudo-true identified set, a profiled likelihood function, and a Rao's score test statistic based on the likelihood's score function.

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

As argued in Aumann (1965), one can view $G(U|x;\theta)$ as the collection of its *measurable selections* (Molchanov and Molinari, 2018, Definition 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 begin with defining the law of $G(U|x;\theta)$ induced by the model's structure:

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

with $\mathcal{C}$ the collection of closed subsets of $\mathcal{Y}$. The functional in (3.1) is the *containment functional* of $G(U|x;\theta)$ and it uniquely determines the distribution of $G(U|x;\theta)$ when it is evaluated at all $A \in \mathcal{C}$ (Molchanov, 2017, p.32).

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 $\tilde{Y} \sim Q(\cdot|x)$,

$$\tilde{Y} \in G(U|x;\theta) \text{ a.s. } \Leftrightarrow \; Q(A|x) \geq \nu_\theta(A|x), \;\; \forall A \subseteq \mathcal{C}, \; x - \text{a.s.}$$

Example 1 (Continued). Let $Y_{x;\theta}(u)$ be the unique element of $G(u|x;\theta)$ if $u \notin M_{x;\theta}$ and $Y_{x;\theta}(u) \equiv (0,0)$ if $u \in M_{x;\theta}$.[10] Molchanov and Molinari (2018, Example 2.6) show that all measurable selections of $G(\cdot|x;\theta)$ in Example 1 can be represented as

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

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* that assigns to $(1,0)$ and $(0,1)$ the probability that each is played given $X = x$ and $U \in M_{x;\theta}$ (e.g., Berry and Tamer, 2006, Ciliberto and Tamer, 2009). Beresteanu et al. (2011, Lemma 2.1) show that the distribution of each selection in Eq. (3.3) belongs to $\text{core}(\nu_\theta(\cdot|x))$, and that only those distributions do. The containment functional of $G(\cdot|x;\theta)$ equals

$$\begin{aligned} \nu_\theta(\{(0,0)\}|x) &= F_\theta(S_{\{(0,0)\}|x;\theta}), && \text{for } A = \{(0,0)\}, \\ \nu_\theta(\{(0,1),(1,0)\}|x) &= 1 - F_\theta(S_{\{(0,0)\}|x;\theta}) - F_\theta(S_{\{(1,1)\}|x;\theta}), && \text{for } A = \{(0,1),(1,0)\}, \end{aligned}$$

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 Eqs. (2.1)- (2.4). □

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 that for all $\theta \in \Theta$, $x \in \mathcal{X}$, and $Q \in \text{core}(\nu_\theta(\cdot|x))$, $Q \ll \mu$.[11] Let the set of conditional densities associated with $\text{core}(\nu_\theta(\cdot|x))$ be

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

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

[10] The assignment for $u \in M_{x;\theta}$ is arbitrary and done only to obtain a random variable defined for all $u \in \mathcal{U}$.

[11] This requirement is typically unrestrictive (see, e.g., White, 1982, p. 2).

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

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

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

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

Define a *model* as the collection of sets $\mathfrak{q}_\theta$ across $\theta \in \Theta$: $\mathfrak{Q} \equiv \{\mathfrak{q}_\theta : \theta \in \Theta\}$. We propose a generalization of the standard definition of correct specification for models with singleton predictions (e.g., White, 1996, Definition 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 Eqs. (3.1) and (3.4)). The model is defined as the collection of (singleton) $q_{\theta,y|x}$ across $\theta \in \Theta$ and $x \in \mathcal{X}$, $\mathcal{Q} = \big\{[q_{\theta,y|x},\, x \in \mathcal{X}] : \; \theta \in \Theta\big\}$. The model is correctly specified if $p_0 = q_\theta$ for some $q_\theta \in \mathcal{Q}$, and misspecified otherwise.*

Given two density functions f and f' on a measure space $(\Omega, \mathfrak{F}, \zeta)$, we measure their similarity through the *Kullback-Leibler Information Criterion (KLIC)*

$$I(f||f') \equiv \int_S f \ln \frac{f}{f'} d\zeta, \quad (3.7)$$

where $S = \{\omega \in \Omega : f(\omega) > 0\}$. In our framework, the model predicts a *set of conditional density functions* as in Eq. (3.5). Hence, we extend the definition in Eq. (3.7) to measure divergence from f of a set of conditional 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* $\int_S f \ln f' d\zeta < \infty$. *The Kullback-Leibler divergence measure from* f *of a set* $\mathfrak{f}$ *is*

$$I(f||\mathfrak{f}) \equiv \inf_{f' \in \mathfrak{f}} I(f||f'). \tag{3.8}$$

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.

Our approach is based on measuring divergence between conditional density functions. Given a joint density function $f(y, x)$, its associated conditional density function $f(y|x)$, and another conditional density function $f'(y|x)$, we denote their conditional KLIC by

$$I(f||f') \equiv \int_{\mathcal{Y} \times \mathcal{X}} f(y, x) \ln \frac{f(y|x)}{f'(y|x)} d\zeta(y, x), \tag{3.9}$$

and use Eq. (3.9) in the KL divergence measure in Eq. (3.8).

Similarly to what White (1982) argued for point-identified models, we define the *pseudo true set*, denoted $\Theta^*(p_0)$, as the set of minimizers of the researcher's ignorance about the true structure. In our case, however, one is ignorant also about which selection from the model predicted set is closest to the data. Hence, minimization occurs not only with respect to $\vartheta \in \Theta$, as in the textbook case for models with singleton predictions, but also with respect to $q \in \mathfrak{q}_\vartheta$. If the model happens to be correctly specified, $\Theta^*(p_0)$ equals the sharp identification region, just like in correctly specified point identified models the pseudo-true value coincides with the data generating one.[12]

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

$$\Theta^*(p_0) \equiv \left\{ \theta \in \Theta : I(p_0||\mathfrak{q}_\theta) = \inf_{\vartheta \in \Theta} I(p_0||\mathfrak{q}_\vartheta) \right\}. \tag{3.10}$$

[12] Chen et al. (2011) provide inference methods for the set in Eq. (3.10) under the assumption of correct model specification. Chen et al. (2018, Remark 3) suggest that their method may remain valid for some misspecified separable models with discrete covariates.

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

$$
\begin{aligned}
I(p_0||\mathfrak{q}_\vartheta) &= \inf_{q \in \mathfrak{q}_\vartheta} \int_{\mathcal{Y}\times\mathcal{X}} p_0(y,x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\zeta(y,x) \\
&= \int_{\mathcal{X}} p_{0,x}(x) \inf_{q_{y|x} \in \mathfrak{q}_{\vartheta,x}} \int_{\mathcal{Y}} p_{0,y|x}(y|x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\mu(y) d\xi(x), \qquad (3.11)
\end{aligned}
$$

where $\mathfrak{q}_\vartheta$ and $\mathfrak{q}_{\vartheta,x}$ are defined in Eqs. (3.4)-(3.5). Eq. (3.11) does not involve unknown selection mechanisms to formalize all ways in which a measurable selection could be picked from $G(\cdot|x;\theta)$ and all associated likelihood functions be obtained. Rather, the optimization step in Eq. (3.11) relies on a convex program, with strictly convex objective and convex constraints (when $\mathcal{Y}$ is finite, it is a finite dimensional convex program with linear constraints), and delivers the density function in $\mathfrak{q}_{\vartheta,x}$ closest to p_0 with respect to KLIC:

$$
q^*_{\vartheta,y|x} = \arg \inf_{q_{y|x} \in \mathfrak{q}_{\vartheta,x}} \int_{\mathcal{Y}} p_{0,y|x}(y|x) \ln \frac{p_{0,y|x}(y|x)}{q_{y|x}(y|x)} d\mu(y). \qquad (3.12)
$$

The solution $q^*_{\vartheta,y|x}$ exists under mild conditions and can be calculated analytically or numerically. 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. Related likelihood-based inference methods, in contrast, rely on an infinite-dimensional parameter space to represent the selection mechanism that picks measurable selections from $G(\cdot|x;\theta)$ (as in, e.g., Eq. (3.3)) and then 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 a discrete distribution (Chen et al., 2018). 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 Example C.1 in Appendix C in the Online Supplement for details).

Putting together Eqs. (3.11) and (3.12) 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 Eq. (3.10) is equal to the set of maximizers of

$$L(\vartheta) \equiv \mathbb{E}_{p_0}\left[\ln q^*_{\vartheta,y|x}(Y|X)\right]. \tag{3.13}$$

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 Eqs. (2.1), (2.2), (2.3), (2.5), let

$$\eta_1(\theta;x) = 1 - F_\theta(S_{\{(0,0)\}|x;\theta}) - F_\theta(S_{\{(1,1)\}|x;\theta}), \tag{3.14}$$

$$\eta_2(\theta;x) = F_\theta(S_{\{(1,0)\}|x;\theta}) + F_\theta(M_{x;\theta}), \tag{3.15}$$

$$\eta_3(\theta;x) = F_\theta(S_{\{(1,0)\}|x;\theta}). \tag{3.16}$$

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

$$\Theta_1(x,p_0) \equiv \left\{\theta \in \Theta : \eta_3(\theta;x) \leq \frac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) \leq \eta_2(\theta;x)\right\} \tag{3.17}$$

$$\Theta_2(x,p_0) \equiv \left\{\theta \in \Theta : \frac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) > \eta_2(\theta;x)\right\} \tag{3.18}$$

$$\Theta_3(x,p_0) \equiv \left\{\theta \in \Theta : \frac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) < \eta_3(\theta;x)\right\}. \tag{3.19}$$

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

$$q^*_{\theta,y|x}((0,0)|x) = F_\theta(S_{\{(0,0)\}|x;\theta}) \tag{3.20}$$

$$q^*_{\theta,y|x}((1,1)|x) = F_\theta(S_{\{(1,1)\}|x;\theta}) \tag{3.21}$$

$$q^*_{\theta,y|x}((0,1)|x) = \begin{cases} \frac{p_{0,y|x}((0,1)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) & \theta \in \Theta_1(x,p_0) \\ \eta_1(\theta;x) - \eta_2(\theta;x) & \theta \in \Theta_2(x,p_0) \\ \eta_1(\theta;x) - \eta_3(\theta;x) & \theta \in \Theta_3(x,p_0) \end{cases} \tag{3.22}$$

$$q^*_{\theta,y|x}((1,0)|x) = \begin{cases} \frac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x) & \theta \in \Theta_1(x,p_0) \\ \eta_2(\theta;x) & \theta \in \Theta_2(x,p_0) \\ \eta_3(\theta;x) & \theta \in \Theta_3(x,p_0) \end{cases} . \tag{3.23}$$

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 $\theta \in \Theta_3(x,p_0)$. □

3.3. *The Score Function*

Here we characterize the *score function* associated with the singleton-valued likelihood function in Eq. (3.13), under the following regularity conditions.

ASSUMPTION 1: *(a) $\mathcal{Y}$ is a finite set. (b) There is a collection $\mathcal{A}_G \subset 2^{\mathcal{Y}}$ such that $\mathcal{A}_G = \text{supp}(G(\cdot|X;\theta)) \equiv \{A \subseteq \mathcal{Y} : F_\theta(G(U|X;\theta) = A) > 0\}$ for all $\theta \in \Theta$, $P_0 - a.s.$ (c) $\nu_\theta(A|X)$ is continuously differentiable with respect to θ for all $A \subset \mathcal{Y}$, $P_0 - a.s.$, and $\nabla_\theta \nu_\theta(A|X)$ is square integrable. (d) $\Theta^*(p_0) \subset \text{int}\,\Theta$. (e) There exists a constant $c > 0$ such that for all $\theta \in \Theta$ and $y \in \mathcal{Y}$, $q^*_{\theta,y|x}(y|x) > c$, $P_0 - a.s.$*

Part (a) of Assumption 1 restricts attention to models with discrete outcomes. Part (b) requires the support of $G(\cdot|X,\theta)$ not to vary with $\theta \in \Theta$.[13] 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 can

[13] Assumption 1-(b) can be weakened to allow $\mathcal{A}_G$ to depend on X, at the cost of heavier notation as the parameter space Θ would depend on X as well. The condition is used in Lemma B.2 in the Online Supplement to show that the Lagrange multiplier vector λ^* associated with the solution of the convex program in Eq. (3.12) is unique. Conditions that imply this uniqueness can replace Assumption 1-(b).

potentially be tested, as shown in Chen and Kaido (2023). Part (c) is easily verified when F_θ is differentiable in θ. Using the first order condition for the likelihood function, part (d) implies that at all $\theta^* \in \Theta^*(p_0)$ the expected score is zero. Part (e) requires $q^*_{\theta,y|x}$ to be bounded away from zero. All conditions are illustrated below, revisiting Example 1.[14]

THEOREM 3.1: *Under Assumption 1, (i)* $L(\theta|X) \equiv \mathbb{E}[\ln q^*_{\theta,y|x}(Y|X)|X]$ *is differentiable with respect to* θ, $P_0 - a.s.$;[15] *(ii) 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*

$$\frac{\partial}{\partial \theta} L(\theta|x) = \mathbb{E}[s_\theta(Y|X; p_{0,y|x})|X = x], \tag{3.24}$$

$$\mathbb{E}[s_\theta(Y|X; p_{0,y|x})] = 0, \ \textit{for all } \theta \in \Theta^*(p_0). \tag{3.25}$$

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 Eq. (3.25) 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 Eq. (3.25). Our Monte Carlo results in Section 5 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. It leverages results in Gauvin and Janin (1990) to establish differentiability with respect to θ of

$$L(\theta|x) = \mathbb{E}[\ln q^*_{\theta,y|x}(Y|X)|X = x] = \sup_{q_{y|x} \in \mathfrak{q}_{\theta,x}} \sum_{y \in \mathcal{Y}} p_{0,y|x}(y|x) \ln q_{y|x}(y|x), \tag{3.26}$$

where $\mathfrak{q}_{\theta,x}$ is defined in Eq. (3.4). The proof also uses results in Ponomarev (2022) and Luo and Wang (2017) that yield that under Assumption 1-(b), the smallest collection of inequalities among the ones in Eq. (3.4) that suffice to sharply characterize $\mathfrak{q}_{\theta,x}$ does not depend on θ. In Lemma B.2 in the Online Supplement, we show that $q^*_{\theta,y|x}$ is insensitive to inclusion

[14] These conditions are also verified for the examples in Appendix C in the Online Supplement.

[15] This result uses only parts (a), (b), (c), and (d) of Assumption 1.

of additional inequalities from Eq. (3.4) in the maximization problem in Eq. (3.26). Hence, so are the pseudo-true set $\Theta^*(p)$, the score function $s_\theta(y|x; p_{0,y|x})$, and the inference procedure that we propose. This is in contrast with much of the 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 (Kédagni et al., 2021) and on the properties of the inference procedure (e.g., Andrews and Soares, 2010, Andrews and Shi, 2013, Bugni et al., 2017, Kaido et al., 2019).

REMARK 3.2: *While our analysis is designed for the use of sharp identifying restrictions through Eq. (3.2), our method remains applicable to a subset of such restrictions, provided the convex program in Eq. (3.26) has a unique solution q^* with unique Lagrange multiplier vector λ^* associated with the constraints (see the proof of Theorem 3.1 part (ii) in Appendix A). This may be relevant in applications where the number of inequalities characterizing the sharp set of model restrictions, even conditional on X, is prohibitively large.*

Example 1 (Continued). Assumption 1-(a) follows immediately. Part (b) holds provided $\delta_1, \delta_2 < \epsilon$ for some $\epsilon < 0$, in which case for all $\theta \in \Theta$, $\mathcal{A}_G = \{\{(0,0)\}, \{(0,1)\}, \{(1,0)\}, \{(1,1)\}, \{(1,0),(0,1)\}\}$.[16] Assumption 1-(c) holds as long as the functions $F_\theta(S_{\{(0,0)\}|x;\theta})$, $F_\theta(S_{\{(1,1)\}|x;\theta})$, $F_\theta(S_{\{(1,0)\}|x;\theta})$ and $F_\theta(M_{x;\theta})$ are differentiable with respect to θ. This can easily be verified for, e.g., the case where U has a bivariate normal distribution with correlation coefficient that is part of θ.[17] 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 Supplement we show that:[18]

$$s_\theta((0,0)|x; p_{0,y|x}) = \frac{\nabla_\theta F_\theta(S_{\{(0,0)\}|x;\theta})}{F_\theta(S_{\{(0,0)\}|x;\theta})} \tag{3.27}$$

[16] If one allows for $\delta_1 = \delta_2 = 0$, at that value the model is complete and part (b) of Assumption 1 does not hold because at that value of θ the support of $G(\cdot|x, \theta)$ changes to $\{\{(0,0)\}, \{(0,1)\}, \{(1,0)\}, \{(1,1)\}\}$.

[17] As the entry game is a threshold crossing model, it is common to assume $Var(U_j) = 1$ for $j = 1, 2$.

[18] Recall that the functions $\eta_1(\theta; x), \eta_2(\theta; x), \eta_3(\theta; x)$ are defined in Eqs. (3.14), (3.15), (3.16). Note also that in this example, the choice probabilities enter $s_\theta(y|x; p_{0,y|x})$ only through the parameter sets $\Theta_1(x; p_{0,y|x})$, $\Theta_2(x; p_{0,y|x})$, $\Theta_3(x; p_{0,y|x})$ defined in Eqs. (3.17), (3.18), (3.19).

$$s_\theta((1,1)|x;p_{0,y|x}) = \frac{\nabla_\theta F_\theta(S_{\{(1,1)\}|x;\theta})}{F_\theta(S_{\{(1,1)\}|x;\theta})}. \tag{3.28}$$

$$s_\theta((0,1)|x;p_{0,y|x}) = \begin{cases} \frac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)} & \theta \in \Theta_1(x;p_{0,y|x}) \\ \frac{\nabla_\theta[\eta_1(\theta;x)-\eta_2(\theta;x)]}{\eta_1(\theta;x)-\eta_2(\theta;x)} & \theta \in \Theta_2(x;p_{0,y|x}) \\ \frac{\nabla_\theta[\eta_1(\theta;x)-\eta_3(\theta;x)]}{\eta_1(\theta;x)-\eta_3(\theta;x)} & \theta \in \Theta_3(x;p_{0,y|x}) \end{cases} \tag{3.29}$$

$$s_\theta((1,0)|x;p_{0,y|x}) = \begin{cases} \frac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)} & \theta \in \Theta_1(x;p_{0,y|x}) \\ \frac{\nabla_\theta \eta_2(\theta;x)}{\eta_2(\theta;x)} & \theta \in \Theta_2(x;p_{0,y|x}) \\ \frac{\nabla_\theta \eta_3(\theta;x)}{\eta_3(\theta;x)} & \theta \in \Theta_3(x;p_{0,y|x}) \end{cases} \tag{3.30}$$

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

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

Any $\theta^* \in \Theta^*(p_0)$ satisfies the population first order condition in Eq. (3.25). Applying the sample analog principle, we propose to estimate $\mathbb{E}[s_{\theta^*}(Y|X;p_{0,y|x})]$ through $\bar{s}_{\theta^*}(\hat{p}_{n,y|x}) \equiv \frac{1}{n}\sum_{i=1}^n s_{\theta^*}(Y_i|X_i;\hat{p}_{n,y|x})$, where $\hat{p}_{n,y|x}$ denotes 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 $p_{0,y|x}$.

In what follows 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

$$\mathbb{G}_{n,\theta}(p)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(s_\theta(Y_i|X_i;p)-\mathbb{E}[s_\theta(Y_i|X_i;p)]\right).$$

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 $\|p_{y|x}-p_{0,y|x}\|_{\mathcal{H}}\leq\delta_n$,

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

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 $\mathbb{E}[s_{\theta^*}(Y,X;p_{y|x})]$. In contrast, Newey (1994) imposes the stronger requirement of smoothness of $s_{\theta^*}(Y,X;p_{y|x})$ with respect to $p_{y|x}$. 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.2-B.3 in the

Online Supplement, 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 3.2: *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 3.2, 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 Eq. (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. 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:

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

where $\hat{\Sigma}_{n,\theta} = \frac{1}{n}\sum_{i=1}^{n}(s_\theta(Y_i|X_i;\hat{p}_{n,y|x}) - \bar{s}_\theta(\hat{p}_{n,y|x}))(s_\theta(Y_i|X_i;\hat{p}_{n,y|x}) - \bar{s}_\theta(\hat{p}_{n,y|x}))^{\top}$ is the sample analog estimator of Σ_θ; $\hat{\Xi}_{n,\theta} = \hat{\Psi}_{n,\theta}^{-1/2}\hat{\Sigma}_{n,\theta}\hat{\Psi}_{n,\theta}^{-1/2}$ is the correlation matrix associated with $\hat{\Sigma}_{n,\theta}$; $\hat{\Psi}_{n,\theta} = \mathrm{diag}(\hat{\Sigma}_{\mathrm{n},\theta})$; and $\varepsilon > 0$ is a regularization constant.

COROLLARY 3.1: *Let Assumptions 1, 2, and 3 hold,* $\min_{j=1,...,d_\theta}\{\text{diag}(\Sigma_{\theta^*})\}_\text{j} > 0$*, and*

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

Then, under $\mathbb{H}_0$ *in Eq. (3.32): (a) For any* $\theta^* \in \Theta^*(p_0)$ *such that* 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.*

Corollary 3.1 requires, in Eq. (3.35), that the population covariance matrix can be consistently estimated; this is a standard requirement in the semiparametric literature.[19] 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 Eq. (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.[20] One can construct a confidence region that covers each point in $\Theta^*(p_0)$ with asymptotic probability $1-\alpha$ as

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

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 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 $O_\mathcal{P}$.

---

[19] In contrast with the semiparametric literature, where Eq. (3.35) is typically required to hold for $\hat{\Sigma}_{n,\hat{\theta}_n}$, with $\hat{\theta}_n$ a consistent estimator of θ^*, in Eq. (3.35) both $\hat{\Sigma}_{n,\theta^*}$ and Σ_{θ^*} are evaluated at the same θ^*.

[20] Nonpivotal asymptotic distributions appear, e.g., in Andrews and Kwon (2022), Andrews and Shi (2013), Andrews and Soares (2010), Kaido et al. (2019) and Bugni et al. (2017). Cox and Shi (2023) propose a test statistic with asymptotically χ^2 distribution, but number of degrees of freedom that depends on θ. On the other hand, Chen et al. (2018) propose test statistics with asymptotically pivotal (χ^2) distribution.

THEOREM 3.3: *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:*[21]

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

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

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 $P \in \mathcal{P}$, $\min_{j=1,\ldots,d_\theta}\{\text{diag}(\Sigma_{\theta^*})\}_j > 0$, *and* $\|\hat{\Sigma}_{n,\theta^*} - \Sigma_{\theta^*}\| = o_{\mathcal{P}}(1)$. *Then, for* CS_n *in Eq. (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 3.3, Corollary 3.1 also applies uniformly over $P \in \mathcal{P}$.

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. 17). 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_τ of $f(\theta) \equiv \ln q^*_{\theta,y|x}$, defined by the convolution:

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

[21] 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.

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

$$\begin{aligned}\frac{\partial}{\partial\theta}f_\tau(\theta) &= \frac{\partial}{\partial\theta}\int f(\theta+\tau z)\phi(z)dz \\ &= -\frac{1}{\tau}\int f(\theta+\tau z)\frac{\partial}{\partial z}\phi(z)dz = -\frac{1}{\tau}\mathbb{E}\left[f(\theta+\tau Z)\frac{\frac{\partial}{\partial z}\phi(Z)}{\phi(Z)}\right],\end{aligned}$$

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_τ. Letting $Z_r, r = 1, \ldots, R$ 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 $\frac{\partial}{\partial\theta}f_\tau(\theta)$ is

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

Replacing f with f_τ introduces a bias proportional to τ, while the variance of $[f(\theta + \tau Z_r) - f(\theta)]Z_r/\tau$ grows with $1/\tau$. In practice one needs to take a stand on this bias-variance trade off. In research-in-progress we formally analyze how to do so. The Monte Carlo approximation in Eq. (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 Eq. (4.1) for each value of (Y_i, X_i). The average score can then be approximated by

$$\frac{1}{n\tau R}\sum_{i=1}^{n}\sum_{r=1}^{R}[f(\theta+\tau Z_{i,r}; Y_i, X_i) - f(\theta; Y_i, X_i)]Z_{i,r}.$$

5. MONTE CARLO EXPERIMENTS

We carry out Monte Carlo simulations based on Example 1, where we denote by X_j a 2×1 vector with first component equal to a constant and second component equal to a

random variable that is either binary or continuously distributed (and player specific). We let $\beta_j = [\beta_{j,1}, \beta_{j,2}]$. The researcher specifies a model where player j earns payoff

$$\pi_j = Y_j(\beta_{j,1} + \beta_{j,2}X_{j,2} + \delta_j Y_{3-j} + U_j),\ Y_j \in \{0,1\}, \tag{5.1}$$

where $U \sim N(0,\Gamma)$, with Γ the 2×2 identity matrix and $\theta = (\delta_1, \delta_2, \beta_{1,1}, \beta_{1,2}, \beta_{2,1}, \beta_{2,2})$.

We consider four data generating processes:

- DGP1: correctly specified model with $X_{j,2} \sim^{i.i.d.} Bernoulli(0.5)$.
- DGP2: correctly specified model with $X_{j,2} \sim^{i.i.d.} N(0,1)$.
- DGP3: misspecified model with $X_{j,2} \sim^{i.i.d.} Bernoulli(0.5)$.
- DGP4: misspecified model with $X_{j,2} \sim^{i.i.d.} N(0,1)$.

We take the true value of θ to be $\theta_0 = (-.7, -.7, .5, .5, .5, .5)$, and generate the data using a selection mechanism where R in Eq. (3.3) is distributed $Bernoulli(0.5)$, independently of all other variables. When the model is correctly specified (DGP1 and DGP2),

$$\begin{aligned}
p_{0,y|x}((0,0)|x) &= q_{\theta_0,y|x}((0,0)|x) = [1-\Phi(x_1\beta_1)][1-\Phi(x_2\beta_2)] \\
p_{0,y|x}((0,1)|x) &= q_{\theta_0,y|x}((0,1)|x) = \eta_1(\theta_0;x) - q_{\theta_0,y|x}((1,0)|x) \\
p_{0,y|x}((1,0)|x) &= q_{\theta_0,y|x}((1,0)|x) = \eta_3(\theta_0;x) + 0.5(\eta_2(\theta_0;x) - \eta_3(\theta_0;x)) \\
p_{0,y|x}((1,1)|x) &= q_{\theta_0,y|x}((1,1)|x) = \Phi(x_1\beta_1+\delta_1)\Phi(x_2\beta_2+\delta_2),
\end{aligned}$$

where the functions $\eta_1(\cdot;x), \eta_2(\cdot;x), \eta_3(\cdot;x)$ are defined in Eqs. (3.14), (3.15), (3.16).

In DGP3 and DGP4, player j's true payoff differs from Eq. (5.1) and is given by

$$\pi_j = Y_j(\beta_{j,1} + \beta_{j,2}X_{j,2} + (\delta_j + \gamma X^*)Y_{3-j} + U_j),\ Y_j \in \{0,1\},$$

where X^* is a binary variable omitted from the model, which affects the strategic interaction effect. Conditional on $(X_{1,2}, X_{2,2}) = (\tilde{x}_1, \tilde{x}_2)$, X^* takes value 1 with probability $\Phi(\frac{\tilde{x}_1-\mu_x}{\sigma_x} + \frac{\tilde{x}_2-\mu_x}{\sigma_x})$, with μ_x and σ_x^2 the mean and variance of $X_{j,2}$. The value of γ determines the extent of misspecification. We report results for $\gamma \in \{-.1, -.2, -.3, -.4, -.5\}$.

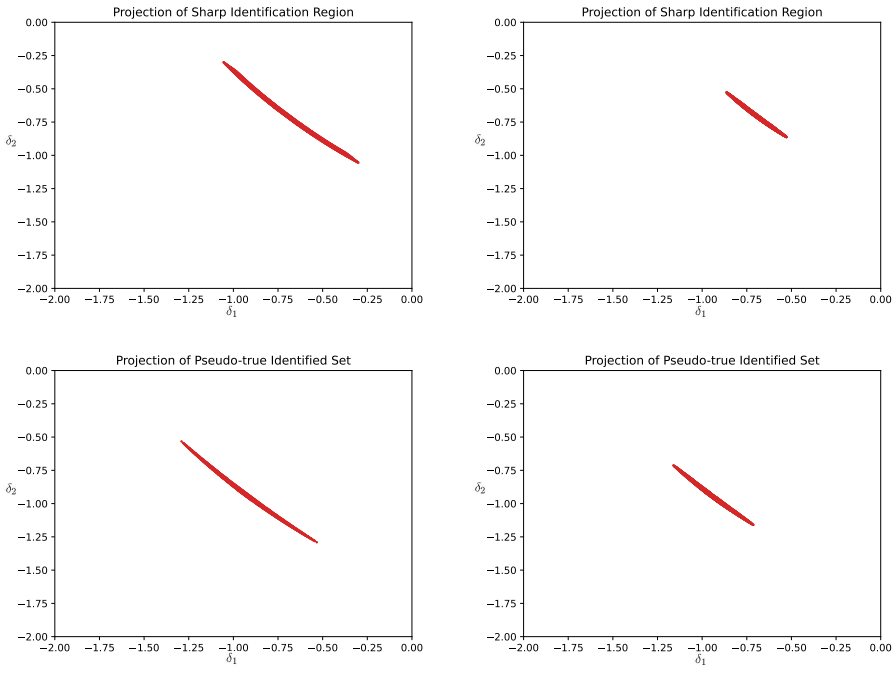


FIGURE 2.—Top: Correctly Specified (left: Design 1 (binary); right: Design 2 (continuous)), Bottom: Misspecified ($\gamma = -.5$; left: Design 3 (binary); right: Design 4 (continuous)).

Figure 2 shows the projections of $\Theta^*(p_0)$ onto the space of (δ_1, δ_2). When the model is correctly specified (top panels), $\Theta^*(p_0)$ coincides with the sharp identification region of θ. With misspecification ($\gamma = -0.5$), the optimal value of the KL divergence measure in Eq. (3.10) is strictly positive (see Table I 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 bottom panels of Figure 2, 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 and slightly stretches out. 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 Eq. (5.1). A reduction in the values of (δ_1, δ_2) (i.e., an increase in absolute value) enlarges the region of multiplicity, thereby allowing for a larger mass to be allocated to $(1, 0)$ and $(0, 1)$ than the model in Eq. (5.1) allows for. As expected, having continuous variation in the covariates substantially reduces the size of $\Theta^*(p_0)$, both in the case of correct specification and with misspecification.

To implement our test, in DGPs 1&3 we estimate $p_{0,y|x}$ using a cell mean estimator. In DGPs 2&4 we use a sieve Logistic estimator with J-th order (tensor-product) Hermite polynomials in (x_{12}, x_{22}) as our sieve space and an L^2 penalty.[22] We compare the performance of our procedure to that of Andrews and Soares (2010) for the case of discrete covariates (DGPs 1&3), and to that of Andrews and Shi (2013) for continuous covariates (DGPs 2&4).[23] We report a power comparison that takes a value θ^* on the boundary of $\Theta^*(p_0)$ and traces rejection probabilities for $\theta_{0,h} = \theta^* + h \times (1, 1, 0, \ldots, 0)'/\sqrt{n}$, with $h > 0$ (hence, we look at drifts of the strategic interaction effects towards $(0, 0)$ that keep the other components fixed) and $n = 2,500$.

Table I reports the results of this exercise for $5,000$ Monte Carlo repetitions. Panel (A) documents size and power of our test as well as the moment inequalities based tests for the case that the model is correctly specified. Both our Rao's score-based test and the tests of Andrews and Soares (2010) and Andrews and Shi (2013) have correct size, although our test slightly underrejects. Nonetheless, both with discrete and with continuous X the power curve of our test quickly dominates that of the moment inequality based tests.

In the misspecified case (Panels (B)-(F)), as expected the tests proposed by Andrews and Soares's (2010) and Andrews and Shi's (2013) are 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 Eq. (3.3)

---

[22] We compute $\tilde{\Sigma}$ in Eq. (3.34) setting $\varepsilon = 0.012$ as recommended by Andrews and Barwick (2012).

[23] We implement the method in Andrews and Soares (2010) using their S_3 test statistic, with moment functions

$$m_{\leq,j}(Z_i, \theta) = \begin{bmatrix} 1\{Y_i = (1,0), X_i = x_j\} - \eta_2(x_j;\theta)p_{x_j} \\ \eta_3(x_j;\theta)p_{x_j} - 1\{Y_i = (1,0), X_i = x_j\} \end{bmatrix}, \tag{5.2}$$

$$m_{=,j}(Z_i, \theta) = \begin{bmatrix} 1\{Y_i = (0,0), X_i = x_j\} - [1 - \Phi(x_1\beta_1)][1 - \Phi(x_2\beta_2)]p_{x_j} \\ 1\{Y_i = (1,1), X_i = x_j\} - \Phi(x_1\beta_1 + \delta_1)\Phi(x_2\beta_2 + \delta_2)\}p_{x_j} \end{bmatrix}, \tag{5.3}$$

$m(z,\theta) = (m_{\leq,1}(z,\theta)', \ldots, m_{\leq,4}(z,\theta)', m_{=,1}(z,\theta)', \ldots, m_{=,4}(z,\theta)')'$, and with $\bar{m}_n(\theta) = \frac{1}{n}\sum_{i=1}^n m(Z_i,\theta)$. We treat p_x as known. This yields a total of 8 inequalities and 8 equalities in the discrete case (where $|\mathcal{X}| = 4$). We implement the method in Andrews and Shi (2013) using their S_2 test stastistic and hyper-cubes as instruments to transform the conditional moment inequalities in unconditional ones. In this case, the number of inequalities is two times the number of hyper-cubes used, and similarly for the equalities.

TABLE I

REJECTION PROBABILITIES OF SCORE AND MOMENT INEQUALITY TESTS

Design	Specification info.	Tests	Size	Power (values of $h/\sqrt{n}$ below)								
				0.021	0.042	0.063	0.084	0.105	0.126	0.147	0.168	0.189
Panel A: Correctly specified ($\gamma = 0$)												
Design 1:	Discrete X											
		Score Test	0.024	0.056	0.195	0.492	0.815	0.966	0.996	1.0	1.0	1.0
		Moment Ineq. Test	0.051	0.077	0.158	0.314	0.535	0.769	0.918	0.983	0.999	1.0
Design 2:	Continuous X											
		Score Test	0.044	0.078	0.206	0.467	0.758	0.941	0.992	1.0	1.0	1.0
		Moment Ineq. Test	0.05	0.081	0.185	0.389	0.688	0.897	0.985	0.999	1.0	1.0
Panel B: Misspecified ($\gamma = -0.1$)												
Design 3:	Discrete X											
	$I(p_0 \| q_\theta^*) = 0.0002$	Score Test	0.036	0.075	0.239	0.554	0.845	0.976	0.998	1.0	1.0	1.0
	IM rej.=0.038	Moment Ineq. Test	0.096	0.12	0.196	0.346	0.557	0.771	0.922	0.985	0.999	1.0
Design 4:	Continuous X											
	$I(p_0 \| q_\theta^*) = 0.0002$	Score Test	0.057	0.092	0.222	0.465	0.754	0.934	0.991	0.999	1.0	1.0
	IM rej.=0.144	Moment Ineq. Test	0.07	0.11	0.221	0.434	0.712	0.905	0.985	0.998	1.0	1.0
Panel C: Misspecified ($\gamma = -0.2$)												
Design 3:	Discrete X											
	$I(p_0 \| q_\theta^*) = 0.0007$	Score Test	0.038	0.079	0.244	0.554	0.848	0.977	0.998	1.0	1.0	1.0
	IM rej.= 0.051	Moment Ineq. Test	0.144	0.174	0.261	0.425	0.621	0.812	0.938	0.988	0.999	1.0
Design 4:	Continuous X											
	$I(p_0 \| q_\theta^*) = 0.0008$	Score Test	0.05	0.081	0.205	0.437	0.724	0.913	0.988	1.0	1.0	1.0
	IM rej.= 0.193	Moment Ineq. Test	0.125	0.179	0.31	0.53	0.771	0.92	0.988	0.999	1.0	1.0
Panel D: Misspecified ($\gamma = -0.3$)												
Design 3:	Discrete X											
	$I(p_0 \| q_\theta^*) = 0.002$	Score Test	0.043	0.084	0.252	0.562	0.851	0.974	0.998	1.0	1.0	1.0
	IM rej. = 0.137	Moment Ineq. Test	0.256	0.306	0.421	0.574	0.748	0.89	0.965	0.994	1.0	1.0
Design 4:	Continuous X											
	$I(p_0 \| q_\theta^*) = 0.002$	Score Test	0.063	0.096	0.225	0.453	0.719	0.910	0.988	0.999	1.0	1.0
	IM rej. = 0.342	Moment Ineq. Test	0.260	0.335	0.486	0.684	0.858	0.953	0.992	1.0	1.0	1.0
Panel E: Misspecified ($\gamma = -0.4$)												
Design 3:	Discrete X											
	$I(p_0 \| q_\theta^*) = 0.003$	Score Test	0.04	0.077	0.228	0.519	0.818	0.965	0.997	1.0	1.0	1.0
	IM rej. = 0.448	Moment Ineq. Test	0.424	0.488	0.598	0.744	0.873	0.953	0.987	0.997	1.0	1.0
Design 4:	Continuous X											
	$I(p_0 \| q_\theta^*) = 0.005$	Score Test	0.067	0.102	0.223	0.447	0.712	0.9	0.98	0.999	1.0	1.0
	IM rej. = 0.634	Moment Ineq. Test	0.465	0.544	0.677	0.826	0.929	0.979	0.997	1.0	1.0	1.0
Panel F: Misspecified ($\gamma = -0.5$)												
Design 3:	Discrete X											
	$I(p_0 \| q_\theta^*) = 0.004$	Score Test	0.042	0.076	0.220	0.501	0.795	0.959	0.996	1.0	1.0	1.0
	IM rej. = 0.817	Moment Ineq. Test	0.640	0.693	0.787	0.879	0.945	0.982	0.996	0.999	1.0	1.0
Design 4:	Continuous X											
	$I(p_0 \| q_\theta^*) = 0.005$	Score Test	0.060	0.096	0.205	0.416	0.677	0.881	0.974	0.998	1.0	1.0
	IM rej. = 0.860	Moment Ineq. Test	0.695	0.756	0.843	0.925	0.972	0.992	0.998	1.0	1.0	1.0

Note: The simulation results are based on random samples of size $n = 2,500$ and 5,000 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 types of covariates (discrete or continuous) and the degree of misspecification measured by the KL divergence, and the rejection probability of an infeasible Information Matrix test.

TABLE II

COMPUTATIONAL TIME COMPARISONS (IN SECONDS)

	Discrete X		Continuous X	
	Score Test	AS10	Score Test	AS13
Calculating the Statistic	0.06	0.01	0.10	2.91
Calculating the Critical Value	0.00	1.08	0.00	166.57

is distributed $Bernoulli(0.5)$. Enriched with this information, the model yields a unique prediction $q_{\theta,y|x}$ 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 I, labeling it "IM rej." As can be seen from the table, for levels of $\gamma \in \{-.1, -.2, -.3\}$ the rejection probability is low, reaching at most 13.7% with discrete covariates and 34.2% with continuous covariates. For $\gamma = -.4, -.5$ the power is higher, reaching 86% in the continuous covariates case. Even for $\gamma = -.3$, the size of the Andrews and Soares (2010) and Andrews and Shi (2013) tests is substantially distorted (about 26% against a 5% nominal level). In contrast, our test has essentially 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 reports average computational time in seconds to calculate test statistics and critical values in DGPs 1 (discrete covariates) and 3 (continuous covariates) for the 5,000 Monte Carlo replications on Boston University's computing cluster (with Intel Xeon Gold 6132 Processors and 192GB RAM). In the discrete case our test statistic takes 0.06 seconds to compute while Andrews and Soares's (2010) test takes 0.01, but our critical value takes 0 seconds to compute as opposed to their 1.08. For the continuous case, our test statistic is twenty nine times faster to compute than Andrews and Shi's (2013), but the most substantial gain comes from calculation of the critical value: zero seconds for us, against 167 for Andrews and Shi (2013).

6. 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." We use their data, but modify their model specification to fully exploit the information provided by the continuously distributed covariates that they discretize as explained below.[24]

Kline and Tamer (2016) 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). They define a market as a trip between two airports, irrespective of intermediate stops. As it is standard in the literature, 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_m^{size} and $X_{j,m}^{pres}$. The first continuously distributed variable is market size, and it enters the payoff of firms of both types in a given market. This variable measures population size at the two endpoints of the trip and is market-specific. The second continuously distributed variable is a firm-and-market-specific variable measuring 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 unobserved payoff shifters $U_{j,m}$ are assumed to have a bivariate normal distribution with $\mathbb{E}(U_{j,m}) = 0$, $Var(U_{j,m}) = 1$, and $Corr(U_{LCC,m}, U_{OA,m}) = r$ for each m and $j \in \{LCC, OA\}$, and to be i.i.d. across m. The correlation parameter r is part of the vector θ and needs to be estimated.[25]

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. We assume that

[24] We downloaded the data from https://www.econometricsociety.org/publications/quantitative-economics/2016/07/01/Bayesian-inference-in-a-class-of-partially-identified-models#supplemental_material.

[25] We assume that $r \in [-0.9, 0.9]$. 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 $c > 0$.

payoffs take the form:

$$\pi_{j,m} = Y_{j,m}(\beta_j^0 + \beta_j^{size} X_m^{size} + \beta_j^{pres} X_{j,m}^{pres} + \delta_j Y_{-j,m} + U_{j,m}). \tag{6.1}$$

In contrast, Kline and Tamer posit

$$\tilde{\pi}_{j,m} = Y_{j,m}(\tilde{\beta}_j^0 + \tilde{\beta}_j^{size} \mathbf{1}(X_m^{size} \geq Med(X^{size})) + \tilde{\beta}_j^{pres} \mathbf{1}(X_{j,m}^{pres} \geq Med(X_j^{pres})) + \tilde{\delta}_j Y_{-j,m} + U_{j,m}). \tag{6.2}$$

In words, they 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. Leveraging our new method, we are able to avoid discretizing the continuously distributed covariates, thereby exploiting all identifying power in their variation.

We analyze 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 $Y_{LCC}(d) = \mathbf{1}(X'_{LCC,m}\beta_{LCC} + \delta_{LCC} d + U_{LCC,m} \geq 0)$, with $\beta_{LCC} = (\beta_{LCC}^0, \beta_{LCC}^{size}, \beta_{LCC}^{pres})$. 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 $P(Y_{LCC}(d) = 1|X_{LCC,m}) = \Phi(X'_{LCC,m}\beta_{LCC} + \delta_{LCC} d)$. We obtain a confidence interval for this parameter for each of $d = 0, 1$ and for specific values of $X_{LCC,m}$. We set X_m^{size} equal to the median of its distribution throughout the analysis. We compute the τ-quantile of the distribution of $X_{LCC,m}^{pres}$ for $\tau \in \mathcal{T} \equiv \{0.125, 0.250, 0.375, 0.5, 0.625, 0.750, 0.875\}$, and then evaluate our parameter of interest for $X_{LCC,m}^{pres}$ set equal to each of these values. Letting $\theta = (\beta_{LCC}^0, \beta_{LCC}^{size}, \beta_{LCC}^{pres}, \delta_{LCC}, \beta_{OA}^0, \beta_{OA}^{size}, \beta_{OA}^{pres}, \delta_{OA}, r)$, we report confidence intervals

$$CI_n(x, d) = \Big[\min_{\theta:\ T_n(\theta) \leq c_{d_\theta, \alpha}} \Phi(x'_{LCC,m}\beta_{LCC} + \delta_{LCC} d), \max_{\theta:\ T_n(\theta) \leq c_{d_\theta, \alpha}} \Phi(x'_{LCC,m}\beta_{LCC} + \delta_{LCC} d) \Big]$$

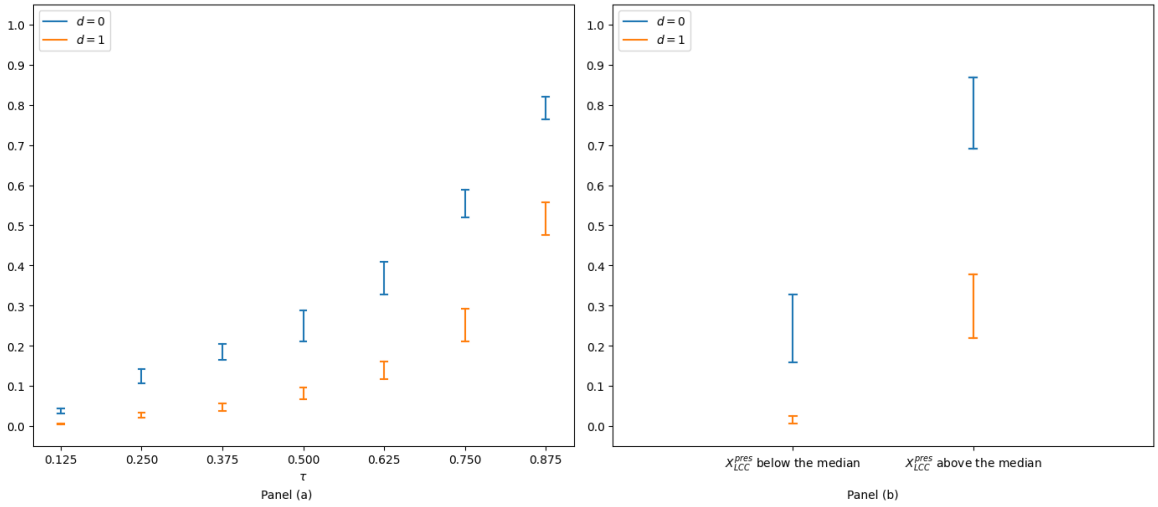


FIGURE 3.—Confidence Intervals for $\Phi(x'_{LCC,m}\beta^*_{LCC} + \delta^*_{LCC}d)$ for $d = 1$ (orange) and $d = 0$ (blue). Panel (a): Rao score test-based inference with $X^{pres}_{LCC,m}$ set equal to the $0.125, 0.250, 0.375, 0.5, 0.625, 0.750, 0.875$ quantiles of its distribution and X^{size}_m set equal to its median. Panel (b): Chen et al. (2018) projection-based inference with $X^{pres}_{LCC,m}$ set equal to 0 if $X^{pres}_{LCC,m}$ is less than or equal to its median, and equal to 1 otherwise.

for the values of x corresponding to $\tau \in \mathcal{T}$ and for $d \in \{0, 1\}$.[26] Under the conditions of Theorem 3.3, by standard arguments

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta^*\in\Theta^*(p)} P(\Phi(x'_{LCC,m}\beta^*_{LCC} + \delta^*_{LCC}d) \in CI_n) \geq 1 - \alpha.$$

Figure 3-Panel (a) reports our results, displaying on the horizontal axis the value of τ and on the vertical axis the candidate value for $\Phi(x'_{LCC,m}\beta^*_{LCC} + \delta^*_{LCC}d)$. The results show substantial heterogeneity in the treatment effects of interest. Entry probabilities are much larger when OA opponents are not in the market (blue segments in Figure 3 for $d = 0$ and orange segments for $d = 1$) across all values of τ, with the effect largest for $\tau \geq 0.750$.

For each fixed value of the entry decision for the OA airlines, as the market presence of the LCC airlines increases, so does the probability that LCC firms enter a market.

[26] Similarly to the Monte Carlo experiments, we estimate $p_{0,y|x}$ using a sieve Logistic estimator with J-th order (tensor-product) Hermite polynomials in $(X^{size}_m, X^{pres}_{LCC,m}, X^{pres}_{OA,m})$ as our sieve space.

However, when OA firms are in the market, the impact of $x^{pres}_{LCC,m}$ on the entry probability is low (the slope of the relationship between entry probability and value of τ is low) until market presence reaches its 0.625 quantile (at which point the slope increases rapidly), suggesting 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 large at all values of τ (the slope is substantial, and it further increases for large values of τ).

We compare our counterfactual estimates of the entry probability for the LCC airlines to what one would obtain using the likelihood based inference method proposed by 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 Eq. (6.2), whereas we use the specification in Eq. (6.1). However, while this implies that the coefficient estimates on $X_{j,m}$ and Y_{-j} are not directly comparable to each other, we believe it to be instructive to compare the counterfactual model-implied entry probabilities.[27]

Figure 3-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. In particular, while it continues to emerge that the presence of OA airlines substantially decreases the probability of entry for LCC airlines, bundling "above the median" and "below the median" as single values for the covariates does not allow one to learn the extent of the nonlinearity in the effect of market presence on the probability of entry. We note that one could use a finer discretization of $(X^{pres}_{LCC,m}, X^{pres}_{OA,m})$ in Eq. (6.2) combined with Chen et al. (2018)'s method. However, doing so would result in a substantially harder computational problem, as 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

[27] We use the replication package provided by Chen et al. (2018) at https://www.econometricsociety.org/publications/econometrica/browse/2018/11/01/monte-carlo-confidence-sets-identified-sets. We adapt the code to yield a confidence interval on $P(Y_{LCC}(d) = 1|X_{LCC,m})$ through the projection method that they propose. Inspection of Chen et al. (2018)'s code shows a slight difference in the payoffs that they specify, $\tilde{\pi}_{j,m} = Y_{j,m}(\tilde{\beta}^0_j + \tilde{\beta}^{size}_j \mathbf{1}(X^{size}_m > Med(X^{size})) + \tilde{\beta}^{pres}_j \mathbf{1}(X^{pres}_{j,m} > Med(X^{pres}_j)) + \tilde{\delta}_j Y_{-j,m} + U_{j,m})$, compared to Eq. (6.2). We compute the confidence intervals using the payoffs in their code and set $\mathbf{1}(X^{size}_m > Med(X^{size})) = 0$.

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 $\mathcal{X}$. Moreover, Chen et al. (2018)'s method requires X to have finite support and hence cannot fully exploit the information provided by continuous variation in X.

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 Eq. (3.26), $L(\theta|x)$ 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 Supplement 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

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

where $\text{core}(\nu_\theta(\cdot|x))$ is defined in Eq. (3.2), and the cardinality of the collection $\mathcal{A}^{(*e)}$ is the smallest among any collection of test sets guaranteeing Eq. (A.1). Hence, it suffices to verify the dominance condition in Eq. (3.2) for all $A \in \mathcal{A}^{(*e)}$ rather than for all $A \in \mathcal{C}$.[28] Second, the problem

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

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 problem 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

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

[28] Galichon and Henry (2011) call collections of sets with this property *core determining*. Applying results in Luo and Wang (2017) and Ponomarev (2022), Lemma B.2 in the Online Supplement shows that $\mathcal{A}^{(*e)}$ is an *exact core determining class*, i.e., it has smallest cardinality among core determining classes.

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.(3.24). In what follows we construct a score function. Let $M = |\mathcal{Y}|$, and order the elements of $\mathcal{Y}$ as $y_1, y_2, \ldots, y_M$. Let $\mathcal{J} = \{j \in \{1, \ldots, J\} : \sum_{\tilde{y} \in A_j} q^*(\tilde{y}) = \nu_\theta(A_j)\}$ be the set of active constraints, and let $\mathcal{J}^c = \{1, \ldots, J\} \setminus \mathcal{J}$ 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

$$\frac{p_0(y_1)}{q^*(y_1)} + \sum_{j \in \mathcal{J}(y_1)} \lambda_j^* = 0 \tag{A.4}$$

$$\vdots$$

$$\frac{p_0(y_M)}{q^*(y_M)} + \sum_{j \in \mathcal{J}(y_M)} \lambda_j^* = 0. \tag{A.5}$$

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)-(A.5) can be written as

$$B\lambda^* = r, \tag{A.6}$$

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

$$B = -\begin{bmatrix} e'_{\mathcal{J}(y_1)} \\ \vdots \\ e'_{\mathcal{J}(y_M)} \end{bmatrix}, \qquad r = \begin{bmatrix} \frac{p_0(y_1)}{q^*(y_1)} \\ \vdots \\ \frac{p_0(y_M)}{q^*(y_M)} \end{bmatrix}. \tag{A.7}$$

By the complementary slackness conditions, $\lambda_j^* = 0$ for any $j \in \mathcal{J}^c$. Hence, (A.6) 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. Eq. (A.6) can be rewritten as

$$\tilde{B}\tilde{\lambda}^* = r, \tag{A.8}$$

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 Eq. (A.8) uniquely, and there exists an S-by-M matrix C such that

$$\tilde{\lambda}^* = Cr. \tag{A.9}$$

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 (A.9). Recall that r is as defined in (A.7). Hence, $\nabla_\theta L(\theta)$ can be written as

$$\nabla_\theta L(\theta) = \sum_{m=1}^M p_0(y_m) \frac{[E_\theta C]_m}{q^*(y_m)}, \tag{A.10}$$

where $[\cdot]_m$ 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 $p_{0,y|x}$. Eq. (A.10) 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)-Eq. (3.25). By law of iterated expectations and dominated convergence theorem,

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

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.*

We next turn to the proof of Theorem 3.2. 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,*

$$D(\theta^*, p_{0,y|x})[p_{y|x} - p_{0,y|x}] = 0. \tag{A.11}$$

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

$$\rho(x,\theta,h_x) = \mathbb{E}_{P_0}[\ln q_{\theta,h_x,y|x}(Y|X)|X=x]. \tag{A.12}$$

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 Eq. (3.12); this selection exists by Artstein (1983). Solving the KL projection problem in Eq. (3.11) (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 $\mathbb{E}_{P_0}[\rho(x,\theta,h(\tau))] \leq \max_{\tilde{h} \in \mathcal{S}(\theta)} \mathbb{E}_{P_0}[\rho(x,\theta,\tilde{h})]$ 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 Eq. (3.24) yields the third equality. Hence, the pathwise derivative is zero. *Q.E.D.*

Proof of Theorem 3.2. Let us write the left-hand side of (3.31) as

$$\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{\theta^*}(Y_i|X_i; \hat{p}_{n,y|x}) = \mathbb{G}_{n,\theta^*}(\hat{p}_{n,y|x}) + \sqrt{n}\mathbb{E}[s_{\theta^*}(Y_i|X_i; \hat{p}_{n,y|x})] \\ &= \mathbb{G}_{n,\theta^*}(p_{0,y|x}) + (\mathbb{G}_{n,\theta^*}(\hat{p}_{n,y|x}) - \mathbb{G}_{n,\theta^*}(p_{0,y|x})) + \sqrt{n}\mathbb{E}[s_{\theta^*}(Y_i|X_i; \hat{p}_{n,y|x})]. \end{aligned} \tag{A.13}$$

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),

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

Hence, by the triangle inequality,

$$|\sqrt{n}\mathbb{E}[s_{\theta^*}(Y_i|X_i; \hat{p}_{n,y|x})]| \leq |\sqrt{n}\mathbb{E}[s_{\theta^*}(Y_i|X_i; p_{0,y|x})]| + |\sqrt{n} r_n| = o_p(1). \tag{A.14}$$

Combining Eqs. (A.13)-(A.14) yields the result in Eq. (3.31). *Q.E.D.*

Proof of Corollary 3.1. For the case that Σ_{θ^*} is nonsingular, standard arguments, Assumption 3, and Eq. (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

$$\zeta = \eta + \nu, \text{ with } \eta \perp\!\!\!\perp \nu, \ \eta \sim N(0, \Sigma_{\theta^*}) \text{ and } \nu \sim N(0, \varepsilon\Psi_{\theta^*}),$$

where Ψ_{θ^*} is the population analog of $\hat{\Psi}_{n,\theta^*}$ (see Eq. (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, 1998, Lemma 8.5),

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

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

Proof of Theorem 3.3. 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 Eq. (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 3.2 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 3.3.

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 3.3.

For Theorem 3.2, under the uniform version of Assumptions 1, 2, and 3 stated in Theorem 3.3, 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 3.2, Eqs. (A.13)-(A.14) 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, 26]

ANDREWS, DONALD W. K. AND PATRIK GUGGENBERGER (2009): "Validity of Subsampling and 'Plug-in Asymptotic' Inference for Parameters Defined by Moment Inequalities," *Econometric Theory*, 25, 669–709. [4]

ANDREWS, DONALD W. K. AND SOONWOO KWON (2022): "Misspecified Moment Inequality Models: Inference and Diagnostics," *Review of Economic Studies*, accepted. [4, 21]

ANDREWS, DONALD W. K. AND XIAOXIA SHI (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666. [5, 17, 21, 26, 28]

ANDREWS, DONALD W. K. AND GUSTAVO SOARES (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157. [4, 5, 17, 21, 26, 28]

ARADILLAS-LOPEZ, ANDRES AND ELIE TAMER (2008): "The Identification Power of Equilibrium in Simple Games," *Journal of Business & Economic Statistics*, 26, 261–283. [8]

ARTSTEIN, Z. (1983): "Distributions of random sets and random selections," *Israel Journal of Mathematics*, 46, 313–324. [3, 9, 37]

AUMANN, ROBERT J (1965): "Integrals of set-valued functions," *Journal of Mathematical Analysis and Applications*, 12, 1–12. [9]

BARSEGHYAN, LEVON, MAURA COUGHLIN, FRANCESCA MOLINARI, AND JOSHUA C. TEITELBAUM (2021): "Heterogeneous Choice Sets and Preferences," *Econometrica*, 89, 2015–2048. [7, 13]

BERESTEANU, ARIE, ILYA MOLCHANOV, AND FRANCESCA MOLINARI (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79, 1785–1821. [3, 8, 10]

BERRY, STEVEN T. AND ELIE TAMER (2006): "Identification in Models of Oligopoly Entry," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by Richard Blundell, Whitney K. Newey, and Torsten Persson, Cambridge University Press, vol. 2 of *Econometric Society Monographs*, 46–85. [10]

BONTEMPS, CHRISTIAN, THIERRY MAGNAC, AND ERIC MAURIN (2012): "Set identified linear models," *Econometrica*, 80, 1129–1155. [4]

BUGNI, FEDERICO A., IVAN A. CANAY, AND PATRIK GUGGENBERGER (2012): "Distortions of Asymptotic Confidence Size in Locally Misspecified Moment Inequality Models," *Econometrica*, 80, 1741–1768. [4]

BUGNI, FEDERICO A., IVAN A. CANAY, AND XIAOXIA SHI (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185, 259 – 282. [4]

——— (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8, 1–38. [5, 17, 21]

CANAY, IVAN A. AND AZEEM M. SHAIKH (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by Bo Honoré, Ariel Pakes, Monika Piazzesi, and Larry Samuelson, Cambridge University Press, vol. 2 of *Econometric Society Monographs*, 271–306. [4]

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. [6, 16]

CHEN, XIAOHONG, TIMOTHY M. CHRISTENSEN, AND ELIE TAMER (2018): "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86, 1965–2018. [6, 12, 13, 21, 31, 32, 33]

CHEN, XIAOHONG, LARS P. HANSEN, AND PETER G. HANSEN (2021): "Robust Inference for Moment Condition Models without Rational Expectations," BF Institute Working Paper N. 2021-122, available at https://bfi.uchicago.edu/wp-content/uploads/2021/10/BFI_WP_2021-122.pdf. [6]

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. [5, 12, 13]

CHRISTENSEN, TIMOTHY AND BENJAMIN CONNAULT (2023): "Counterfactual Sensitivity and Robustness," *Econometrica*, 91, 263–298. [6]

CILIBERTO, FEDERICO AND ELIE TAMER (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77, 1791–1828. [5, 8, 10]

COX, GREGORY AND XIAOXIA SHI (2023): "Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models," *The Review of Economic Studies*, 90, 201–228. [21]

DICKSTEIN, MICHAEL J AND EDUARDO MORALES (2018): "What do Exporters Know?" *The Quarterly Journal of Economics*, 133, 1753–1801. [5]

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. [22]

GALICHON, ALFRED AND MARC HENRY (2009): "A test of non-identifying restrictions and confidence regions for partially identified parameters," *Journal of Econometrics*, 152, 186 – 196. [4]

——— (2011): "Set Identification in Models with Multiple Equilibria," *The Review of Economic Studies*, 78, 1264–1298. [34]

GALLANT, A. RONALD AND HALBERT WHITE (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, B. Blackwell. [2]

GAUVIN, JACQUES AND ROBERT JANIN (1990): "Directional derivative of the value function in parametric optimization," *Annals of Operations Research*, 27, 237–252. [16, 34]

GUGGENBERGER, PATRIK, JINYONG HAHN, AND KYOOIL KIM (2008): "Specification testing under moment inequalities," *Economics Letters*, 99, 375 – 378. [4]

HALL, ALASTAIR R. AND ATSUSHI INOUE (2003): "The large sample behaviour of the generalized method of moments estimator in misspecified models," *Journal of Econometrics*, 114, 361 – 394. [2]

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

HECKMAN, JAMES J. (1978): "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence," *Annales de l'inséé*, 227–269. [7]

HONORÉ, BO E. AND ELIE TAMER (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629. [7]

KAIDO, HIROAKI, FRANCESCA MOLINARI, AND JÖRG STOYE (2019): "Confidence Intervals for Projections of Partially Identified Parameters," *Econometrica*, 87, 1397–1432. [5, 17, 21]

KAIDO, HIROAKI AND HALBERT WHITE (2013): "Estimating Misspecified Moment Inequality Models," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*, ed. by Xiaohong Chen and Norman R. Swanson, Springer, New York, NY, 331–361. [4]

KAIDO, HIROAKI AND YI ZHANG (2019): "Robust Likelihood Ratio Tests for Incomplete Economic Models," available at https://arxiv.org/abs/1910.04610. [6]

KÉDAGNI, DÉSIRÉ, LIXIONG LI, AND ISMAËL MOURIFIÉ (2021): "Discordant Relaxations of Misspecified Models," available at https://arxiv.org/abs/2012.11679. [3, 5, 17]

KLINE, BRENDAN AND ELIE TAMER (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7, 329–366. [5, 6, 29, 30]

LUO, YE AND HAI WANG (2017): "Core Determining Class and Inequality Selection," *AER P&P*, 107, 274–77. [16, 34]

MAGNOLFI, LORENZO AND CAMILLA RONCORONI (2023): "Estimation of Discrete Games with Weak Assumptions on Information," *The Review of Economic Studies*, 90, 2006–2041. [8]

MANSKI, CHARLES F. (2003): *Partial Identification of Probability Distributions*, Springer Series in Statistics, Springer. [4]

MOLCHANOV, I. (2017): *Theory of Random Sets*, London: Springer, 2 ed. [9]

MOLCHANOV, ILYA AND FRANCESCA MOLINARI (2018): *Random Sets in Econometrics*, Econometric Society Monograph Series, Cambridge University Press, Cambridge UK. [5, 7, 9, 10]

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, 3, 4, 7]

NEWEY, WHITNEY K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382. [3, 18, 19, 37]

DE PAULA, ÁUREO, SETH RICHARDS-SHUBIK, AND ELIE TAMER (2018): "Identifying Preferences in Networks With Bounded Degree," *Econometrica*, 86, 263–288. [8]

PONOMAREV, KIRILL (2022): "Essays in Econometrics," Ph.D. thesis, UCLA. [16, 34]

PONOMAREVA, MARIA AND ELIE TAMER (2011): "Misspecification in moment inequality models: back to moment equalities?" *The Econometrics Journal*, 14, 186–203. [4]

ROCKAFELLAR, R. T. (1984): *Directional differentiability of the optimal value function in a nonlinear programming problem*, Berlin, Heidelberg: Springer Berlin Heidelberg, 213–226. [34]

ROMANO, JOSEPH P. AND AZEEM M. SHAIKH (2008): "Inference for identifiable parameters in partially identified econometric models," *Journal of Statistical Planning and Inference*, 138, 2786 – 2807. [4]

SHENG, SHUYANG (2020): "A Structural Econometric Analysis of Network Formation Games Through Subnetworks," *Econometrica*, 88, 1829–1858. [8]

STOYE, JÖRG (2020): "A Simple, Short, but Never-Empty Confidence Interval for Partially Identified Parameters," available at https://arxiv.org/abs/2010.10484. [5]

TAMER, ELIE (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70, 147–165. [7, 8]

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press. [39]

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

WHITE, HALBERT (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25. [2, 3, 10, 12, 26, 33]

——— (1996): *Estimation, inference and specification analysis*, 22, Cambridge university press. [11, 12]

ONLINE APPENDIX FOR: "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

APPENDIX B: LEMMAS USED IN PROOFS OF MAIN THEOREMS

B.1. *Pseudo-True Sets and Score Function*

We give three lemmas that we use to show the differentiability of $L(\theta|x)$. To ease notation, we drop conditioning on X. We let $\mathcal{C}$ denote the collection of closed subsets of $\mathcal{Y}$. Molchanov and Molinari (2018, Section 2.2) show that $\text{core}(\nu_\theta(\cdot))$ can be expressed as $\text{core}(\nu_\theta(\cdot)) \equiv \left\{Q \in \mathcal{M}(\Sigma_Y) : \nu_\theta(A) \leq Q(A) \leq \nu^*_\theta(A), A \subseteq \mathcal{C}\right\}$, with $\nu^*_\theta(A) \equiv \int_{\mathcal{U}} \mathbf{1}(G(u;\theta) \cap A \neq \emptyset) dF_\theta(u)$.[1] Let $\mathcal{A} \subseteq 2^{\mathcal{Y}}$ be a collection of events. Among sets in $\mathcal{A}$, let $\mathcal{A}_=$ collect all restrictions such that $\nu_\theta(A) = \nu^*_\theta(A)$. That is, the sets belonging to $\mathcal{A}_=$ imply equality restrictions. We then let $\mathcal{A}_\geq$ collect the remaining events. Let Δ denote the

Hiroaki Kaido: hkaido@bu.edu

Francesca Molinari: fm72@cornell.edu

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

[1] For a given compact set $A \subseteq \mathcal{Y}$, $\nu^*_\theta(A)$ is the *capacity functional* of $G(u;\theta)$ (Molchanov and Molinari, 2018, Definition 1.23), and $\nu_\theta(A) = 1 - \nu^*_\theta(A^c)$, with A^c the complement of A and $\nu^*_\theta(\cdot)$ extended to the family of open sets (as in, e.g., Molchanov, 2017, Theorem 1.1.21), and where ν_θ is the containment functional of $G(\cdot;\theta)$.

$|\mathcal{Y}|-1$ dimensional unit-simplex. Consider the following problem:

$$\mathbf{P}(\mathcal{A}): \qquad \upsilon(\theta;\mathcal{A}) \equiv \max_{q\in\Delta} \sum_{y\in\mathcal{Y}} p_0(y)\ln q(y) \tag{B.1}$$

$$s.t. \quad \sum_{y\in A} q(y) \geq \nu_\theta(A),\ A\in\mathcal{A}_{\geq} \tag{B.2}$$

$$\sum_{y\in A} q(y) = \nu_\theta(A),\ A\in\mathcal{A}_{=}, \tag{B.3}$$

Let $A\subset\mathcal{Y}$. As $\mathcal{Y}$ is finite, one may represent the probability that any distribution P with probability mass function $p\in\Delta$ assigns to a set A through a *representer* a of A, by writing

$$P(A) = p^\top a,$$

with $a\in\{0,1\}^{|\mathcal{Y}|}$. For example, take $\mathcal{Y}=\{(0,0),(0,1),(1,0),(1,1)\}$, $A=\{(1,0),(1,1)\}$, and $a=(0,0,1,1)^\top$. Then, $P(A)=p^\top a$. Similarly, for some $b_A\in\{0,1\}^{|\mathcal{Y}|}$, the constraints in (B.2)-(B.3) can be written as

$$q^\top b_A \geq \nu_\theta(A),\ A\in\mathcal{A}_{\geq}$$
$$q^\top b_A = \nu_\theta(A),\ A\in\mathcal{A}_{=}.$$

For any pair of sets $A_1, A_2\subset\mathcal{Y}$ with associated representer vectors $a^1, a^2\in\{0,1\}^{|\mathcal{Y}|}$, their union $A_1\cup A_2$ and intersections $A_1\cap A_2$ are represented by $a^1\vee a^2$ (componentwise maximum) and $a^1\wedge a^2$ (componentwise minimum). For any event $A\subseteq\mathcal{Y}$ with $A=\cup_{i=1}^k A_i$,

$$P(A) = \sum_{I\neq\emptyset, I\subseteq\{1,\ldots,k\}} (-1)^{|I|+1} P\Big(\bigcap_{i\in I} A_i\Big).$$

In terms of corresponding vectors,

$$p^\top a = \sum_{I\neq\emptyset, I\subseteq\{1,\ldots,k\}} (-1)^{|I|+1} p^\top \Big(\bigwedge_{i\in I} a^i\Big).$$

Since this holds for any p in the probability simplex, we must have

$$a = \sum_{I \neq \emptyset, I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} (\bigwedge_{i \in I} a^i). \tag{B.4}$$

This means that the representers of the events A and $A_1, \ldots, A_k$ are linearly dependent. The following lemma shows that the opposite is also true.

LEMMA B.1: *Let $a^0, \ldots, a^k$ be the representers of $A_0, \ldots, A_k$. The following statements are equivalent*

1. *$\{ \bigwedge_{i \in I} a^i, I \neq \emptyset, I \subseteq \{0, \ldots, k\} \}$ are linearly dependent.*
2. *There exists $j \in \{0, \ldots, k\}$ such that $A_j = \bigcup_{i \in \{0,\ldots,k\} \setminus \{j\}} A_i$.*

PROOF: We let the elements of $\mathcal{Y}$ be ordered as $\{y_1, y_2, \ldots, y_{|\mathcal{Y}|}\}$. (2. $\Rightarrow$ 1.) follows immediately from (B.4). For (1. $\Rightarrow$ 2.), w.l.o.g. take $j = 0$ and suppose that $C \equiv A_0 \setminus \cup_{i=1}^k A_i$ is nonempty. Let $I_C \subset \{1, \ldots, |\mathcal{Y}|\}$ collect the indices of outcomes belonging to C. Take $y_j \in C \subset \mathcal{Y}$. Then, the representer e_j of y_j is a $|\mathcal{Y}|$-dimensional vector whose j-th component is 1, and the remaining components are all 0s. Note that $y_j \notin \cup_{i=1}^k A_i$ implies that the j-th component of $\bigwedge_{i \in I} a^i$ is 0 for any $I \neq \emptyset, I \subseteq \{1, \ldots, k\}$. Hence, e_j cannot be expressed as a linear combination of $\{ \bigwedge_{i \in I} a^i, I \neq \emptyset, I \subseteq \{1, \ldots, k\} \}$. Since $y_j \in A_0$, this in turn means, a^0 cannot be expressed as a linear combination of $\{ \bigwedge_{i \in I} a^i, I \neq \emptyset, I \subseteq \{1, \ldots, k\} \}$. Hence, $\{a^0, \bigwedge_{i \in I} a^i, I \neq \emptyset, I \subseteq \{1, \ldots, k\} \}$ are linearly independent. The case with $\cup_{i=1}^k A_i \setminus A_0 \neq \emptyset$ can be analyzed similarly. *Q.E.D.*

LEMMA B.2: *Let Assumption 1 hold. Then, (i) there exists a collection of subsets of $\mathcal{Y}$ denoted $\mathcal{A}^{(*e)}$ that does not depend on $\theta \in \Theta$ such that* $\text{core}(\nu_\theta(\cdot)) = \{Q \in \mathcal{M}(\Sigma_Y) : Q(A) \geq \nu_\theta(A), A \in \mathcal{A}^{(*e)}\}$, *with* $\text{core}(\nu_\theta(\cdot))$ *defined in Eq. (3.2), and no collection of sets $\mathcal{A}^*$ of cardinality smaller than $\mathcal{A}^{(*e)}$ suffices to characterize* $\text{core}(\nu_\theta(\cdot))$*; (ii) the optimal value of $\mathbf{P}(\mathcal{A}^{(*e)})$ is $L(\theta|x)$; (iii) the solution $q^* \in \Delta$ to the problem $\mathbf{P}(\mathcal{A}^{(*e)})$ is unique, and it also solves problem $\mathbf{P}(\mathcal{A}^{all})$, with $\mathcal{A}^{all} = \{A : A \subseteq \mathcal{Y},\ A$ closed$\}$; (iv) the associated Lagrange multiplier vector λ^* is unique.*

PROOF: **Part (i).** A family of closed sets $\mathcal{A}^*$ is a *core determining class* (Galichon and Henry, 2011) if any probability measure Q defined on $\mathcal{Y}$ satisfying the inequalities $Q(A) \geq \nu_\theta(A)$ for all $A \in \mathcal{A}^*$ satisfies the inequalities $Q(A) \geq \nu_\theta(A)$ for all closed sets $A \subseteq \mathcal{Y}$. A family of closed sets $\mathcal{A}$ is an *exact* core determining class (Luo and Wang, 2017) if it has the smallest cardinality among all core determining classes. Using results in Telgen (1983), Ponomarev (2022, Eqs. (2.3)-(2.4), p.81) shows that when $\mathcal{Y}$ is a finite set, the exact core determining class for $G(\cdot;\theta)$ equals

$$\begin{aligned} \mathcal{A}^{(*e)} &= \{A \subseteq \mathcal{Y} : q^M(A) < \nu_\theta(A)\}, \\ \text{where} \quad q^M(A) &= \min_{q \in \Delta} \left\{ q^\top b_A : \; q^\top b_{\tilde{A}} \geq \nu_\theta(\tilde{A}) \text{ for all } \tilde{A} \subseteq \mathcal{Y}, \tilde{A} \neq A \right\} \end{aligned}$$

By Assumption 1, $\mathcal{Y}$ is finite and the (finite) support $\mathcal{A}_G$ of the correspondence $G(\cdot|\theta)$ is fixed for all $\theta \in \Theta$, P_0-a.s. Due to the finiteness of $\mathcal{Y}$, arguing as in Ponomarev (2022, Theorem 2.2), one can build a partition of $\mathcal{U}$, denoted $\mathcal{U}_\theta = \{u_1, \ldots, u_\kappa\}$ corresponding to the set of values of U associated with each realization $\bar{G}_k \in \mathcal{A}_G$, $k = 1, \ldots, \kappa$. As $\mathcal{A}_G$ is constant across all values of $\theta \in \Theta$, so is the bipartite graph from $\mathcal{U}_\theta$ to $\mathcal{A}_G$, even though $\mathcal{U}_\theta$ (and the probability that F_θ assigns to each element of the partition) does change with θ. Ponomarev (2022, Theorem 2.2) shows that $\mathcal{A}^{(*e)}$ can be characterized using exclusively the connectedness properties of the subgraphs induced by $(A, \{u \in \mathcal{U}_\theta : \bar{G}(u) \subseteq A\})$ and $(A^c, \{u \in \mathcal{U}_\theta : \bar{G}(u) \cup A^c \neq \emptyset\})$. As the bipartite graph from $\mathcal{U}_\theta$ to $\mathcal{A}_G$ is constant across θ, the connectedness structure of the subgraphs induced by $(A, \{u \in \mathcal{U}_\theta : \bar{G}(u) \subseteq A\})$ and $(A^c, \{u \in \mathcal{U}_\theta : \bar{G}(u) \cup A^c \neq \emptyset\})$ is also constant across $\theta \in \Theta$, and hence so is $\mathcal{A}^{(*e)}$.

Part (ii). By the definition of an exact core determining class, the collection of inequalities in $A^{(*e)}$ yields the same constraint set as in Eq. (3.26). The solution to $\mathbf{P}(\mathcal{A}^{(*e)})$ exists by the continuity of the objective function and the compactness of the probability simplex.

Part (iii). As $q \mapsto \mathbb{E} \ln q$ is strictly concave and the domain of q is convex, uniqueness of q^* follows. As the collection of inequalities in $A^{(*e)}$ yields the same constraint set as in Eq. (3.26), q^* solves also the original problem in Eq. (3.26).

Part (iv). The constraint set consists of linear (in)equalities. Hence, the Karush-Kuhn-Tucker conditions hold at the feasible point q^* with Lagrange multiplier λ^*. A sufficient condition for λ^* to be unique is that the Linear Independence Constraint Qualification (LICQ) holds. To establish this, we first note that the full set of constraints can be expressed as (e.g., Molchanov and Molinari, 2018, Section 2.2):

$$\nu_\theta^*(A) \geq q^\top b_A \geq \nu_\theta(A),\ A \subset \mathcal{Y} : 1 \leq |A| \leq \lceil |\mathcal{Y}|/2 \rceil, \tag{B.5}$$

where $|\cdot|$ denotes the cardinality of the set in its argument, and $\lceil \cdot \rceil$ represents the smallest integer greater than or equal to its argument. Eq. (B.5) follows because for any set $A \subset \mathcal{Y}$ and its complement $A^c = \mathcal{Y} \setminus A$, one has $b_A = 1 - b_{A^c}$ and $\nu_\theta(A) = 1 - \nu_\theta^*(A^c)$. This implies that the gradient of any pair of inequalities in (B.5) equals a representer b_A. Moreover, either only one of the two inequalities in (B.5) can be an *active* inequality, or we are in the presence of an equality restriction. Next, recall that we are further restricting the collection of sets in (B.5) to be the ones in $\mathcal{A}^{(*e)}$. If the LICQ condition fails, there must exist a collection of sets $A_j \in \{A \subset \mathcal{A}^{(*e)} : 1 \leq |A| \leq \lceil |\mathcal{Y}|/2 \rceil\},\ j = 1, \ldots, k$, such that $q^\top b_{A_j} = \nu_\theta(A_j)$, $q^\top b_A = \nu_\theta(A)$, and, by Lemma B.1, $A = \cup_{j=1,\ldots,k} A_j$. This in turn implies

$$\nu_\theta(A) \geq \sum_{I \neq \emptyset, I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} \nu_\theta\Big(\bigcap_{i \in I} A_i\Big),$$

and we have that the inequality for the set A is satisfied whenever the inequalities for the sets $A_j, j = 1, \ldots, k$ are satisfied. But this contradicts $\mathcal{A}^{(*e)}$ being an exact core determining class because such a set A could be removed from it. *Q.E.D.*

B.2. *Derivation of Results and Verification of Conditions for the Entry Game Example 1*

B.2.1. *Profiled likelihood and score function*

PROPOSITION B.1: *Under the assumptions laid out in Example 1, (i) the profiled likelihood $q^*_{\theta,y|x}$ for $y \in \{(0,0),(0,1),(1,0),(1,1)\}$ is given in equations (3.20)-(3.23). (ii) The score function is given in equations (3.27)-(3.30).*

PROOF: In this example $q_{y|x}((0,1)|x) = \eta_1(\theta;x) - q_{y|x}((1,0)|x)$. We use directly in the optimization problem the fact that the elements of the vector $q_{y|x}(\cdot|x)$ sum to one. Let $z \equiv q_{y|x}((1,0)|x)$, so that $q_{y|x}((0,1)|x) = \eta_1(\theta;x) - z$ and $c(\theta) = p_{0,y|x}((0,0)|x) \ln F_\theta(S_{\{(0,0)\}|x;\theta}) + p_{0,y|x}((1,1)|x) \ln F_\theta(S_{\{(1,1)\}|x;\theta})$. Rewrite the optimization problem as

$$V(\theta) = \sup_z c(\theta) + p_{0,y|x}((1,0)|x) \ln z + p_{0,y|x}((0,1)|x) \ln(\eta_1(\theta;x) - z) \tag{B.6}$$

$$s.t.\ z - \eta_3(\theta;x) \geq 0 \tag{B.7}$$

$$\eta_2(\theta;x) - z \geq 0, \tag{B.8}$$

Define the Lagrangian of this problem by

$$\mathcal{L}(z,\lambda,\theta) = c(\theta) + p_{0,y|x}((1,0)|x) \ln z + p_{0,y|x}((0,1)|x) \ln(\eta_1(\theta;x) - z) + \lambda_1(z - \eta_3(\theta;x)) + \lambda_2(\eta_2(\theta;x) - z).$$

Part (i). Since c does not affect the solution, we drop it in the analysis that follows. The Karush-Kuhn-Tucker (KKT) conditions of the problem under study are

$$-p_{0,y|x}(1,0|x)\tfrac{1}{z} + p_{0,y|x}(0,1|x)\tfrac{1}{\eta_1(\theta;x)-z} - \lambda_1 + \lambda_2 = 0 \tag{B.9}$$

$$\lambda_1(\eta_3(\theta;x) - z) = 0 \tag{B.10}$$

$$\lambda_2(z - \eta_2(\theta;x)) = 0 \tag{B.11}$$

$$\lambda_1, \lambda_2 \geq 0. \tag{B.12}$$

Below, we consider three cases.

Case 1: ($\lambda_1 = \lambda_2 = 0$.) If $\lambda_1 = \lambda_2 = 0$, solving (B.9) yields

$$z = \frac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x) + p_{0,y|x}((0,1)|x)} \eta_1(\theta;x). \tag{B.13}$$

Therefore, the resulting distribution $q^*_{\theta,y|x}$ is given by

$$(q^*_{\theta,y|x}((0,0)|x), q^*_{\theta,y|x}((1,1)|x), q^*_{\theta,y|x}((0,1)|x), q^*_{\theta,y|x}((1,0)|x))^\top = \Big(F_\theta(S_{\{(0,0)\}|x;\theta}), F_\theta(S_{\{(1,1)\}|x;\theta}), \frac{p_{0,y|x}((0,1)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x), \frac{p_{0,y|x}((1,0)|x)}{p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x)}\eta_1(\theta;x)\Big)^\top$$

For z in (B.13) to be an optimal solution, it needs to satisfy the inequality restrictions in (B.7)-(B.8), which is true iff θ belongs to $\Theta_1(x)$.

Case 2: ($\lambda_2 > 0$.) If $\lambda_2 > 0$, the complementary slackness condition (B.11) implies $z = \eta_2(\theta;x)$, hence $q^*_{\theta,y|x}((1,0)|x)$ equals its upper bound. The minimizing density is then

$$(q^*_{\theta,y|x}((0,0)|x), q^*_{\theta,y|x}((1,1)|x), q^*_{\theta,y|x}((0,1)|x), q^*_{\theta,y|x}((1,0)|x))^\top = \Big(F_\theta(S_{\{(0,0)\}|x;\theta}), F_\theta(S_{\{(1,1)\}|x;\theta}), \eta_1(\theta;x) - \eta_2(\theta;x), \eta_2(\theta;x)\Big)^\top.$$

For this to be an optimal solution, one needs to ensure that $\lambda_2 > 0$. Substituting $z = \eta_2(\theta;x)$ into (B.9) and solving for λ_2 yields

$$\lambda_2 = p_{0,y|x}((1,0)|x)\frac{1}{\eta_2(\theta;x)} - p_{0,y|x}((0,1)|x)\frac{1}{\eta_1(\theta;x)-\eta_2(\theta;x)}$$

It is then straightforward to show that $\lambda_2 > 0$ iff θ belongs to $\Theta_2(x)$.

Case 3: ($\lambda_1 > 0$.) The result follows using the same argument as in Case 2.

Part (ii). To establish this result, we let $\theta(t) = \theta + th, h \in \mathbb{R}^d$ and define $\tilde{V}(t) = V(\theta(t))$, $\tilde{\mathcal{L}}(z,\lambda,t) = \mathcal{L}(z,\lambda,\theta(t))$. so that the Lagrangian becomes

$$\tilde{\mathcal{L}}_t(z,\lambda,\theta(t)) = p_{0,y|x}((0,0)|x)\frac{\nabla_\theta F_\theta(S_{\{(0,0)\}|x;\theta})^\top h}{F_\theta(S_{\{(0,0)\}|x;\theta})} + p_{0,y|x}((1,1)|x)\frac{\nabla_\theta F_\theta(S_{\{(1,1)\}|x;\theta})^\top h}{F_\theta(S_{\{(1,1)\}|x;\theta})} + p_{0,y|x}((0,1)|x)\frac{\nabla_\theta \eta_1(\theta;x)^\top h}{\eta_1(\theta;x)-z} + \lambda_1 \nabla_\theta \eta_3(\theta;x)^\top h + \lambda_2 \nabla_\theta \eta_2(\theta;x)^\top h.$$

Therefore, for any sequence $\{t_n\}$ with $t_n \downarrow 0$, the maximizer of $\mathcal{L}(z, \lambda, \theta + t_n h)$ exists. The domain of the control variable and parameter $[0,1] \times (-\epsilon, \epsilon)$ is bounded, and hence the inf-boundedness assumption of Rockafellar (1984) holds. This ensures that the parametric optimization problem indexed by t is directionally stable in the sense of Gauvin and Janin (1990). Using the fact that (q^*, λ^*) are unique as shown above, we can apply Gauvin and Janin (1990, Corollary 4.2) to obtain full differentiability of V. The derivative of $\tilde{V}(t)$ can be derived analytically as follows.

Case 1: $\lambda_1 = \lambda_2 = 0$

$$\tilde{V}'(t) = p_{0,y|x}((0,0)|x) \frac{\nabla_\theta F_\theta(S_{\{(0,0)\}|x;\theta})^\top h}{F_\theta(S_{\{(0,0)\}|x;\theta})} + p_{0,y|x}((1,1)|x) \frac{\nabla_\theta F_\theta(S_{\{(1,1)\}|x;\theta})^\top h}{F_\theta(S_{\{(1,1)\}|x;\theta})} + \left[p_{0,y|x}((1,0)|x) + p_{0,y|x}((0,1)|x)\right] \frac{\nabla_\theta \eta_1(\theta;x)^\top h}{\eta_1(\theta;x)}.$$

Case 2: $\lambda_1 = 0, \lambda_2 > 0$

$$\tilde{V}'(t) = p_{0,y|x}((0,0)|x) \frac{\nabla_\theta F_\theta(S_{\{(0,0)\}|x;\theta})^\top h}{F_\theta(S_{\{(0,0)\}|x;\theta})} + p_{0,y|x}((1,1)|x) \frac{\nabla_\theta F_\theta(S_{\{(1,1)\}|x;\theta})^\top h}{F_\theta(S_{\{(1,1)\}|x;\theta})} + p_{0,y|x}((1,0)|x) \frac{\nabla_\theta \eta_2(\theta;x)^\top h}{\eta_2(\theta;x)} + p_{0,y|x}((0,1)|x) \frac{\nabla_\theta [\eta_1(\theta;x) - \eta_2(\theta;x)]^\top h}{\eta_1(\theta;x) - \eta_2(\theta;x)}.$$

Case 3: $\lambda_1 > 0, \lambda_2 = 0$

$$\tilde{V}'(t) = p_{0,y|x}((0,0)|x) \frac{\nabla_\theta F_\theta(S_{\{(0,0)\}|x;\theta})^\top h}{F_\theta(S_{\{(0,0)\}|x;\theta})} + p_{0,y|x}((1,1)|x) \frac{\nabla_\theta F_\theta(S_{\{(1,1)\}|x;\theta})^\top h}{F_\theta(S_{\{(1,1)\}|x;\theta})} + p_{0,y|x}((1,0)|x) \frac{\nabla_\theta \eta_3(\theta;x)^\top h}{\eta_3(\theta;x)} + p_{0,y|x}((0,1)|x) \frac{\nabla_\theta [\eta_1(\theta;x) - \eta_3(\theta;x)]^\top h}{\eta_1(\theta;x) - \eta_3(\theta;x)}.$$

Comparing the expressions for these three cases with equations (3.27)-(3.30), we obtain $\frac{\partial}{\partial \theta} \mathbb{E}[\ln q^*_{\theta,y|x}(Y|X)|X = x] = \mathbb{E}[s_\theta(Y|X)|X = x]$. *Q.E.D.*

B.2.2. *Verification of Assumptions*

Assumptions 2, 3 (iv), and the requirement in Eq. (3.35), are high-level conditions that need to be verified in each application of our method. Below we do so for the en-

try game example, under some regularity conditions. We first provide some notation that will be useful throughout. Let $\|p_{y|x}\|_{\mathcal{H}} = \sup_{y\in\mathcal{Y}} \sup_{x\in\mathcal{X}} |p_{y|x}(y|x)|$. We use the functions $\eta_1(\theta;x), \eta_2(\theta;x), \eta_3(\theta;x)$ defined in Eqs. (3.14)-(3.16). For $j = 1, 2$, we let:

$$Z_j(X;p_{y|x}) \equiv p_{y|x}((1,0)|X)\eta_1(\theta;X) - (p_{y|x}((1,0)|X) + p_{y|x}((0,1)|X))\eta_{j+1}(\theta;X), \quad \text{(B.14)}$$

and note that $Z_1(x;p_{y|x}) \le Z_2(x;p_{y|x})$ because $\eta_2(\theta;X) \ge \eta_3(\theta;X)$. We use the functions Z_1, Z_2 to define indicators $\mathbb{I}_\ell$ that re-express the sets $\Theta_\ell, \ell = 1, 2, 3$, in Eqs. (3.17)-(3.19):

$$\begin{aligned}
\mathbb{I}_1(x;p_{y|x}) &= 1\big\{p_{y|x}((1,0)|x)\eta_1(\theta;x) - (p_{y|x}((1,0)|x) + p_{y|x}((0,1)|x))\eta_2(\theta;x) \le 0\big\} \\
&\quad \times 1\big\{p_{y|x}((1,0)|x)\eta_1(\theta;x) - (p_{y|x}((1,0)|x) + p_{y|x}((0,1)|x))\eta_3(\theta;x) \ge 0\big\} \\
&= 1\big\{Z_1(x;p_{y|x}) \le 0\big\}1\{Z_2(x;p_{y|x}) \ge 0\} \qquad \text{(B.15)} \\
\mathbb{I}_2(x;p_{y|x}) &= 1\big\{p_{y|x}((1,0)|x)\eta_1(\theta;x) - (p_{y|x}((1,0)|x) + p_{y|x}((0,1)|x))\eta_2(\theta;x) > 0\big\} \\
&= 1\{Z_1(x;p_{y|x}) > 0\} \qquad \text{(B.16)} \\
\mathbb{I}_3(x;p_{y|x}) &= 1\big\{p_{y|x}((1,0)|x)\eta_1(\theta;x) - (p_{y|x}((1,0)|x) + p_{y|x}((0,1)|x))\eta_3(\theta;x) < 0\big\} \\
&= 1\{Z_2(x;p_{y|x}) < 0\}. \qquad \text{(B.17)}
\end{aligned}$$

One may rewrite the score functions in Eqs. (3.27)-(3.30) as

$$\begin{aligned}
s_\theta((0,0)|x;p_{y|x}) &= \tfrac{\nabla_\theta F_\theta(S_{\{(0,0)\}|x;\theta})}{F_\theta(S_{\{(0,0)\}|x;\theta})} \\
s_\theta((0,1)|x;p_{y|x}) &= \tfrac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)}\mathbb{I}_1(x;p_{y|x}) + \tfrac{\nabla_\theta[\eta_1(\theta;x)-\eta_2(\theta;x)]}{\eta_1(\theta;x)-\eta_2(\theta;x)}\mathbb{I}_2(x;p_{y|x}) \\
&\quad + \tfrac{\nabla_\theta[\eta_1(\theta;x)-\eta_3(\theta;x)]}{\eta_1(\theta;x)-\eta_3(\theta;x)}\mathbb{I}_3(x;p_{y|x}) \\
s_\theta((1,0)|x;p_{y|x}) &= \tfrac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)}\mathbb{I}_1(x;p_{y|x}) + \tfrac{\nabla_\theta \eta_2(\theta;x)}{\eta_2(\theta;x)}\mathbb{I}_2(x;p_{y|x}) \\
&\quad + \tfrac{\nabla_\theta \eta_3(\theta;x)}{\eta_3(\theta;x)}\mathbb{I}_3(x;p_{y|x}) \\
s_\theta((1,1)|x;p_{y|x}) &= \tfrac{\nabla_\theta F_\theta(S_{\{(1,1)\}|x;\theta})}{F_\theta(S_{\{(1,1)\}|x;\theta})}
\end{aligned}$$

For any vector $a = (a_1, \ldots, a_{d_X})$, define the differential operator by

$$D^{|a|} = \frac{\partial^{|a|}}{\partial x_1^{a_1} \cdots \partial x_d^{a_{d_X}}},$$

where $|a| = \sum_i^{d_X} a_i$. Then, for a function $h : \mathcal{X} \to \mathbb{R}$, let

$$\|h\|_{\infty,\alpha} = \max_{|a| \leq [\alpha]} \sup_x |D^a h(x)| + \max_{|a|=[\alpha]} \sup_{x \neq x'} \frac{|D^a h(x) - D^a h(x')|}{\|x - x'\|^{\alpha - [\alpha]}}.$$

Let $\mathcal{C}_M^\alpha(\mathcal{X})$ be the set of continuous functions $h : \mathcal{X} \to \mathbb{R}$ with $\|h\|_{\infty,\alpha} \leq M$.

We next provide regularity conditions under which we verify Assumptions 2 and 3 (iv) .

ASSUMPTION B.1: *For the entry game model in Example 1,*

(i) There exists $C > 0$ s.t. $\|\nabla_\theta \eta_j(\theta; x)\| \leq C,\ j = 1, \ldots, 3$, for all $x \in \mathcal{X}$.

(ii) There exists $c > 0$ s.t. $\mathcal{H} = \{p_{y|x} : \mathcal{X} \to [0,1]^{\mathcal{Y}} : p_{y|x}(y|x) \geq c, \forall (y,x) \in \mathcal{Y} \times \mathcal{X}\}$.

(iii) If X has at least one component with continuous distribution, the probability density function of $Z_j|X_d$, for $j = 1, 2$, is uniformly bounded on the support of $Z_j|X_d$, where X_d denotes the subvector of X containing discrete covariates with finite support. If there are no discrete covariates, the restriction is on the unconditional probability density function of Z_j.

ASSUMPTION B.2:

(a) One of the following conditions hold:

(i) X is a vector of discrete random variables and $\mathcal{X} \subset \mathbb{R}^{d_X}$ is a finite set.

(ii) X is a vector of continuous random variables and $\mathcal{X} \subset \mathbb{R}^{d_X}$ is a bounded, convex set with nonempty interior. For some $c > 0$, $M > 0$, and $\alpha > d_X$,

$$\mathcal{H} = \{p_{y|x} : \mathcal{X} \to [0,1]^{|\mathcal{Y}|} : p_{y|x}(y|\cdot) \in \mathcal{C}_M^\alpha(\mathcal{X}), y \in \mathcal{Y},\ p_{y|x}(y|x) \geq c > 0, \forall (y,x) \in \mathcal{Y} \times \mathcal{X}\}$$

(iii) $X = (X_c^\top, X_d^\top)^\top$ *consists of subvectors* X_c *and* X_d*, where* X_c *is continuously distributed and* X_d *is discretely distributed.* $\mathcal{X} = \mathcal{X}_c \times \mathcal{X}_d \subset \mathbb{R}^{d_X}$*, where* $\mathcal{X}_c \subset \mathbb{R}^{d_{X_c}}$ *is a bounded convex set with nonempty interior, and* $\mathcal{X}_d \subset \mathbb{R}^{d_{X_d}}$ *is a finite set. For some* $c > 0$*,* $M > 0$*,* $\alpha > d_X$*, Lipschitz functions* $\phi_k, k = 1, \ldots, |\mathcal{Y}|$*, and some functions* ℓ_c *and* ℓ_d*,*

$$\mathcal{H} = \{p_{y|x} : \mathcal{X} \to [0,1]^{|\mathcal{Y}|} : p_{y|x}(y_k|x) = \phi_k\big(\ell_c(y_k|x_c), \ell_d(y_k|x_d)\big),\ \ell_c(y_k|\cdot) \in \mathcal{C}_M^\alpha(\mathcal{X}_c),$$
$$-M \leq \ell_d(y_k|x_d) \leq M, \forall x_d \in \mathcal{X}_d,\ k = 1, \ldots, d_Y,\ p_{y|x}(y|x) \geq c > 0, \forall (y,x) \in \mathcal{Y} \times \mathcal{X}\}.$$

(b) $\mathbb{E}[\|\frac{\nabla_\theta F_\theta(S_{\{y\}|x;\theta})}{F_\theta(S_{\{y\}|x;\theta})}\|^2] \leq C$ *for* $y = (0,0), (1,1)$ *for some* $0 < C < \infty$.

REMARK B.1: Assumption B.1(i) is satisfied, for example, when the vector U has a multivariate Normal distribution, provided the correlation among any two of its entries is bounded away from one (in absolute value). In Assumption B.2(a)(iii), we assume that $p_{y|x}$ combines a function of continuous covariates X_c with a function of discrete covariates X_d using a Lipschitz function, which covers many transformations of interest (see, e.g., van der Vaart and Wellner, 1996, p. 192). More general transformations can be allowed for, as far as one may ensure that the metric entropy of $\mathcal{H}$ can be controlled properly.

PROPOSITION B.2: *Suppose Assumptions 1 and B.1 hold for the entry game model in Example 1. Then Assumption 2 also holds.*

PROOF: Recall that $m_\theta(x; p_{y|x}) \equiv \mathbb{E}[s_\theta(Y|X; p_{y|x})|X = x] = \sum_{y \in \mathcal{Y}} p_{0,y|x}(y|x) s_\theta(y, x; p_{y|x})$. For $p_{y|x}, p_{0,y|x} \in \mathcal{H}$, our goal is to bound $\mathbb{E}[\|m_\theta(X; p_{y|x}) - m_\theta(X; p_{0,y|x})\|]$. Observe that the score depends on the underlying conditional probability only through $\mathbb{I}(x; p_{y|x}) = (\mathbb{I}_1(x; p_{y|x}), \mathbb{I}_2(x; p_{y|x}), \mathbb{I}_3(x; p_{y|x}))$. Hence, the difference $\Delta(x; p_{y|x}, p_{0,y|x}) \equiv \|m_\theta(x; p_{y|x}) - m_\theta(x; p_{0,y|x})\|$ can only be nonzero if $\mathbb{I}(x; p_{y|x}) \neq \mathbb{I}(x; p_{0,y|x})$. The exact values of the difference is summarized in Table B.I. Below, we consider two subcases (i) X is discrete and $\mathcal{X}$ is finite, and (ii) X contains a continuously distributed variable. The case in which all components of X are continuously distributed is treated as a special case of the latter.

TABLE B.I

VALUES OF $\Delta(x; p_{y|x}, p_{0,y|x})$ WHEN $\mathbb{I}(x; p_{y|x}) \neq \mathbb{I}(x; p_{0,y|x})$

$\mathbb{I}(x; p_{y\|x})$	$\mathbb{I}(x; p_{0,y\|x})$	$\Delta(x; p_{y\|x}, p_{0,y\|x})$
$(1,0,0)$	$(0,1,0)$	$\Big\|p_{0,y\|x}((1,0)\|x)\Big(\frac{\nabla_\theta \eta_2(\theta;x)}{\eta_2(\theta;x)} - \frac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)}\Big)$
$(0,1,0)$	$(1,0,0)$	$+ p_{0,y\|x}((0,1)\|x)\Big(\frac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_2(\theta;x)}{\eta_1(\theta;x) - \eta_2(\theta;x)} - \frac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)}\Big)\Big\|$
$(1,0,0)$	$(0,0,1)$	$\Big\|p_{0,y\|x}((1,0)\|x)\Big(\frac{\nabla_\theta \eta_3(\theta;x)}{\eta_3(\theta;x)} - \frac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)}\Big)$
$(0,0,1)$	$(1,0,0)$	$+ p_{0,y\|x}((0,1)\|x)\Big(\frac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_3(\theta;x)}{\eta_1(\theta;x) - \eta_3(\theta;x)} - \frac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)}\Big)\Big\|$
$(0,1,0)$	$(0,0,1)$	$\Big\|p_{0,y\|x}((1,0)\|x)\Big(\frac{\nabla_\theta \eta_2(\theta;x)}{\eta_2(\theta;x)} - \frac{\nabla_\theta \eta_3(\theta;x)}{\eta_3(\theta;x)}\Big)$
$(0,0,1)$	$(0,1,0)$	$+ p_{0,y\|x}((0,1)\|x)\Big(\frac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_2(\theta;x)}{\eta_1(\theta;x) - \eta_2(\theta;x)} - \frac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_3(\theta;x)}{\eta_1(\theta;x) - \eta_3(\theta;x)}\Big)\Big\|$

(i) Discrete X: Let $\mathcal{X}$ be a finite set and let $\mathcal{X}_0 = \{x \in \mathcal{X} : Z_1(x; p_{0,y|x}) \neq 0, Z_2(x; p_{0,y|x}) \neq 0\}$. Let $c \equiv \min_{x \in \mathcal{X}_0} \min_{j=1,2} |Z_j(x; p_{0,y|x})|$. Then, by Lemma B.3, $\Delta(x; p_{y|x}, p_{0,y|x}) = 0$ for all $x \in \mathcal{X}_0$ and $p_{y|x}$ such that $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta$ for all $\delta \leq c/4$. Hence, they do not contribute to the L^1-norm of $\Delta(\cdot; p_{y|x}, p_{0,y|x})$. Now consider $x \in \mathcal{X}_0^c$. For example, suppose that $0 = Z_1(x; p_{0,y|x}) < Z_2(x; p_{0,y|x})$. To make $\Delta(x; p_{y|x}, p_{0,y|x})$ nonzero, let $p_{y|x}$ be such that $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta$ and $Z_1(x; p_{y|x}) > 0$. This leads to $\mathbb{I}(x, p_{y|x}) = (0,1,0)$ and $\mathbb{I}(x, p_{0,y|x}) = (1,0,0)$. From Table B.I,

$$\Delta(x; p_{y|x}, p_{0,y|x}) = \Big\| p_{0,y|x}((1,0)|X) \frac{\nabla_\theta \eta_2(\theta;X)}{\eta_2(\theta;X)} + p_{0,y|x}((0,1)|X) \frac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_2(\theta;x)}{\eta_1(\theta;x) - \eta_2(\theta;x)} - [p_{0,y|x}((1,0)|X) + p_{0,y|x}((0,1)|X)] \frac{\nabla_\theta \eta_1(\theta;X)}{\eta_1(\theta;X)} \Big\|. \quad \text{(B.18)}$$

Note that $Z_1(x; p_{0,y|x}) = 0$ is equivalent to

$$p_{0,y|x}((1,0)|X) = [p_{0,y|x}((1,0)|X) + p_{0,y|x}((0,1)|X)] \frac{\eta_2(\theta;X)}{\eta_1(\theta;X)}. \quad \text{(B.19)}$$

Since $\frac{p_{0,y|x}((0,1)|X)}{p_{0,y|x}((1,0)|X)+p_{0,y|x}((0,1)|X)} = 1 - \frac{p_{0,y|x}((1,0)|X)}{p_{0,y|x}((1,0)|X)+p_{0,y|x}((0,1)|X)}$, we obtain

$$p_{0,y|x}((0,1)|X) = [p_{0,y|x}((1,0)|X) + p_{0,y|x}((0,1)|X)]\frac{\eta_1(\theta;X)-\eta_2(\theta;X)}{\eta_1(\theta;X)}. \tag{B.20}$$

Substituting (B.19)-(B.20) into (B.18) yields,

$$\Delta(x; p_{y|x}, p_{0,y|x}) = \Bigg\| [p_{0,y|x}((1,0)|X) + p_{0,y|x}((0,1)|X)] \Big(\frac{\nabla_\theta \eta_2(\theta;X)}{\eta_1(\theta;X)} + \frac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_2(\theta;x)}{\eta_1(\theta;x)} - \frac{\nabla_\theta \eta_1(\theta;X)}{\eta_1(\theta;X)} \Big) \Bigg\| = 0.$$

A similar argument can be applied to $x \in \mathcal{X}_0^c$ such that $Z_1(x; p_{0,y|x}) < Z_2(x; p_{0,y|x}) = 0$. Finally, consider $x \in \mathcal{X}_0^c$ such that $Z_1(x; p_{0,y|x}) = Z_2(x; p_{0,y|x}) = 0$. This occurs only if $\eta_2(\theta; x) = \eta_3(\theta; x)$. Hence, $Z_1(x; p_{y|x}) = Z_2(x; p_{y|x})$ for any $p_{y|x}$. It then suffices to consider only one of Z_j's. For example let $p_{y|x}$ be such that $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta$ and $Z_1(x; p_{y|x}) > 0$. Then, the same analysis as above leads to $\Delta(x; p_{y|x}, p_{0,y|x}) = 0$.

Therefore, for all $p_{y|x}$ such that $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta$ for a sufficiently small δ,

$$\begin{aligned} \Big\| \mathbb{E}\Big[m_\theta(X; p'_{y|x}) - m_\theta(X; p_{0,y|x}) \Big] \Big\| &\leq \mathbb{E}\Big[\| m_\theta(X; p'_{y|x}) - m_\theta(X; p_{0,y|x}) \| \Big] \\ &= \sum_{x \in \mathcal{X}_0} p_{0,x}(x) \Delta(x; p_{y|x}, p_{0,y|x}) + \sum_{x \in \mathcal{X}_0^c} p_{0,x}(x) \Delta(x; p_{y|x}, p_{0,y|x}) = 0 \end{aligned}$$

Therefore, the pathwise derivative is 0.

(ii) X contains a continuously distributed variable:

Let X_d be a subvector of X containing discrete covariates. Recall that $\Delta(x, p_{y|x}, p_{0,y|x}) \neq 0$ when $\mathbb{I}(x; p_{y|x}) \neq \mathbb{I}(x; p_{0,y|x})$. This occurs when $\operatorname{sgn}(Z_j(x; p_{y|x})) \neq \operatorname{sgn}(Z_j(x; p_{0,y|x}))$ for some j. By Eq. (B.14) and $\sup_{x \in \mathcal{X}} |\eta_j(\theta; x)| \leq 1$, $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta$ implies

$$\sup_{x \in \mathcal{X}} |Z_j(x; p_{y|x}) - Z_j(x; p_{0,y|x})| \leq 3\delta, \ j = 1, 2$$

Therefore, if $\text{sgn}(Z_j(x;p_{y|x})) \neq \text{sgn}(Z_j(x;p_{0,y|x}))$ and $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta$, one must have $|Z_j(x;p_{0,y|x})| \leq 3\delta$. Hence,

$$\begin{aligned}
&\Big\| E\Big[m_\theta(X;p_{y|x}) - m_\theta(X;p_{0,y|x})\Big] \Big\| \leq \mathbb{E}\Big[\|m_\theta(X;p_{y|x}) - m_\theta(X;p_{0,y|x})\|\Big] \\
&\quad \leq \mathbb{E}\Big[\Delta(X;p_{y|x},p_{0,y|x})(1\{-3\delta \leq Z_1(X;p_{0,y|x}) \leq 3\delta\} + 1\{-3\delta \leq Z_2(X;p_{0,y|x}) \leq 3\delta\})\Big] \\
&\quad \leq K\delta\mathbb{E}\Big[\int 1\{-3\delta \leq z_1 \leq 3\delta\} f_{Z_1|X_d}(z_1)dz_1 + \int 1\{-3\delta \leq z_2 \leq 3\delta\} f_{Z_2|X_d}(z_2)dz_2\Big] \\
&\quad \leq c\delta^2
\end{aligned}$$

where the third line follows by Lemma B.4 and the law of iterated expectations, the fourth line uses Assumption B.1 (iii), and $0 < c < \infty$ is some constant. Therefore, the pathwise derivative is again zero. *Q.E.D.*

LEMMA B.3: *Let $\mathcal{X}$ be a finite set, and let $\mathcal{X}_0 = \{x \in \mathcal{X} : Z_1(x;p_{0,y|x}) \neq 0, Z_2(x;p_{0,y|x}) \neq 0\}$. Let $c \equiv \min_{x\in\mathcal{X}_0} \min_{j=1,2} |Z_j(x;p_{0,y|x})|$. Then, $\Delta(x;p_{y|x},p_{0,y|x}) = 0$ for any $x \in \mathcal{X}_0$ and $p_{y|x}$ such that $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq c/4$.*

PROOF: Take $x \in \mathcal{X}_0$. Suppose $Z_1(x;p_{0,y|x}) \geq c > 0$ so that $\mathbb{I}(x;p_{0,y|x}) = (0,1,0)$. Let $p_{y|x}$ satisfy $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq c/4$. Then, by (B.14) and $|\eta_j(x;\theta)| \leq 1$, and the triangle inequality, $Z_1(x;p_{y|x}) \geq Z_1(x;p_{0,y|x}) - \frac{3}{4}c \geq \frac{1}{4}c > 0$, implying $\mathbb{I}(x;p_{y|x}) = (0,1,0)$. From Table B.I, $\Delta(x;p_{y|x},p_{0,y|x}) = 0$. Other cases can be analyzed similarly. *Q.E.D.*

LEMMA B.4: *Suppose Assumptions 1 and B.1 hold for the entry game model in Example 1. For $\delta > 0$, let $p_{y|x}$ be such that $\|p_{y|x} - p_{0,y|x}\|_{\mathcal{H}} \leq \delta$. Then, there exists $0 < K < \infty$ such that for all $x \in \mathcal{X}$, $\Delta(x;p_{y|x},p_{0,y|x}) \leq K\delta$.*

PROOF: From Table B.I, $\Delta(x;p_{y|x},p_{0,y|x}) = 0$ when $\mathbb{I}(x;p_{y|x}) = \mathbb{I}(x;p_{0,y|x})$. Therefore, we focus on cases with $\mathbb{I}(x;p_{y|x}) \neq \mathbb{I}(x;p_{0,y|x})$ below. Consider the case where $\mathbb{I}(x;p_{y|x}) = (0,1,0)$ and $\mathbb{I}(x;p_{0,y|x}) = (1,0,0)$. By (B.15)-(B.17), this occurs when

$$Z_1(x;p_{0,y|x}) \leq 0, \;\; Z_1(x;p_{y|x}) > 0. \tag{B.21}$$

Furthermore, by (B.14) and $\sup_{x\in\mathcal{X}}|\eta_j(\theta;x)|\leq 1$, $\|p_{y|x}-p_{0,y|x}\|_{\mathcal{H}}\leq\delta$ implies

$$\sup_{x\in\mathcal{X}}|Z_1(x;p_{y|x})-Z_1(x;p_{0,y|x})|\leq 3\delta. \tag{B.22}$$

Combining (B.21)-(B.22) yields $-3\delta\leq Z_1(x;p_{0,y|x})\leq 0$. By (B.14) and $\eta_j(\theta;x)\geq c$,

$$\begin{aligned}&(p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x))\tfrac{\eta_2(\theta;x)}{\eta_1(\theta;x)}-\tfrac{3}{c}\delta\\&\qquad\leq p_{0,y|x}((1,0)|x)\leq(p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x))\tfrac{\eta_2(\theta;x)}{\eta_1(\theta;x)}\end{aligned} \tag{B.23}$$

Using Assumption B.1 (ii), this may also be written as

$$\tfrac{\eta_2(\theta;x)}{\eta_1(\theta;x)}-\tfrac{3}{c(p_{0,y|x}((1,0)|X)+p_{0,y|x}((0,1)|X))}\delta\leq\tfrac{p_{0,y|x}((1,0)|x)}{(p_{0,y|x}((1,0)|X)+p_{0,y|x}((0,1)|X))}\leq\tfrac{\eta_2(\theta;x)}{\eta_1(\theta;x)}.$$

Since $\frac{p_{0,y|x}((0,1)|x)}{(p_{0,y|x}((1,0)|X)+p_{0,y|x}((0,1)|X))}=1-\frac{p_{0,y|x}((1,0)|x)}{(p_{0,y|x}((1,0)|X)+p_{0,y|x}((0,1)|X))}$, we obtain

$$1-\tfrac{\eta_2(\theta;x)}{\eta_1(\theta;x)}\leq\tfrac{p_{0,y|x}((0,1)|x)}{(p_{0,y|x}((1,0)|X)+p_{0,y|x}((0,1)|X))}\leq 1-\tfrac{\eta_2(\theta;x)}{\eta_1(\theta;x)}+\tfrac{3}{c(p_{0,y|x}((1,0)|X)+p_{0,y|x}((0,1)|X))}\delta.$$

This may in turn be written as

$$\begin{aligned}&(p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x))\tfrac{\eta_1(\theta;x)-\eta_2(\theta;x)}{\eta_1(\theta;x)}\\&\qquad\leq p_{0,y|x}((0,1)|x)\leq(p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x))\tfrac{\eta_1(\theta;x)-\eta_2(\theta;x)}{\eta_1(\theta;x)}+\tfrac{3}{c}\delta.\end{aligned} \tag{B.24}$$

By (B.23) and (B.24), let us write

$$p_{0,y|x}((1,0)|x)=(p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x))\tfrac{\eta_2(\theta;x)}{\eta_1(\theta;x)}+r_{(1,0)}(x) \tag{B.25}$$

$$p_{0,y|x}((0,1)|x)=(p_{0,y|x}((1,0)|x)+p_{0,y|x}((0,1)|x))\tfrac{\eta_1(\theta;x)-\eta_2(\theta;x)}{\eta_1(\theta;x)}+r_{(0,1)}(x), \tag{B.26}$$

where $r_{(1,0)}(x) \in [-3\delta/c, 0]$ and $r_{(0,1)}(x) \in [0, 3\delta/c]$ for all $x \in \mathcal{X}$. From Table B.I, the value of $\Delta(x; p_{y|x}, p_{0,y|x})$ when $\mathbb{I}(x; p_{y|x}) = (0,1,0)$ and $\mathbb{I}(x; p_{0,y|x}) = (1,0,0)$ is

$$\Delta(x; p_{y|x}, p_{0,y|x}) = \Big\| p_{0,y|x}((1,0)|x) \tfrac{\nabla_\theta \eta_2(\theta;x)}{\eta_2(\theta;x)} + p_{0,y|x}((0,1)|x) \tfrac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_2(\theta;x)}{\eta_1(\theta;x) - \eta_2(\theta;x)} - [p_{0,y|x}((1,0)|x) + p_{0,y|x}((0,1)|x)] \tfrac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)} \Big\|. \quad \text{(B.27)}$$

By (B.25)-(B.26), the terms inside the norm in (B.27) can therefore be written as

$$[p_{0,y|x}((1,0)|x) + p_{0,y|x}((0,1)|x)] \left(\tfrac{\nabla_\theta \eta_2(\theta;x)}{\eta_1(\theta;x)} + \tfrac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_2(\theta;x)}{\eta_1(\theta;x)} - \tfrac{\nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)} \right) + \tfrac{\nabla_\theta \eta_2(\theta;x)}{\eta_2(\theta;x)} r_{(1,0)}(x) + \tfrac{\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_2(\theta;x)}{\eta_1(\theta;x) - \eta_2(\theta;x)} r_{(0,1)}(x).$$

By the triangle inequality, $\eta_j(\theta; x) \geq c$, and Assumption B.1 (i), we obtain

$$\begin{aligned} \Delta(x; p_{y|x}, p_{0,y|x}) &\leq \|\nabla_\theta \eta_2(\theta; x)\| \left| \tfrac{r_{(1,0)}(x)}{\eta_2(\theta;x)} \right| + (\|\nabla_\theta \eta_1(\theta; x)\| + \|\nabla_\theta \eta_2(\theta; x)\|) \left| \tfrac{r_{(0,1)}(x)}{\eta_1(\theta;x) - \eta_2(\theta;x)} \right| \\ &\leq \tfrac{3C}{c^2}\delta + \tfrac{6C}{c^2}\delta, \end{aligned}$$

which establishes the claim of the lemma for $\mathbb{I}(x; p_{y|x}) = (0,1,0)$ and $\mathbb{I}(x; p_{0,y|x}) = (1,0,0)$. The other cases can be analyzed similarly. *Q.E.D.*

We next establish stochastic equicontinuity for the empirical process

$$\mathbb{G}_n(p) = \tfrac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(s_\theta(Y_i|X_i; p) - \mathbb{E}[s_\theta(Y_i|X_i; p)] \right), \quad p \in \mathcal{H}$$

In the definition of $\mathbb{G}_n$, the structural parameter θ is fixed. Hence, the function class $\mathcal{F} = \{f(y,x) : f(y,x) = s_\theta(y|x; p), p \in \mathcal{H}\}$ is defined by mixing and matching p with fixed functions such as $\eta_j(x, \theta)$. For example, we may write $s_\theta((1,0)|x; p)$ as

$$s_\theta((1,0)|x; p_{0,y|x}) = \sum_{j=1}^{3} \tfrac{\nabla_\theta \eta_j(\theta;x)}{\eta_j(\theta;x)} \mathbb{I}_j(x; p_{y|x}),$$

where $\mathbb{I}_j(x; p_{y|x}), j = 1, 2, 3$, are defined in Eqs. (B.15)-(B.17).

PROPOSITION B.3: *Suppose Assumptions 1, B.1, and B.2 hold for the entry game model in Example 1. Then Assumption 3 (iv) also holds.*

PROOF: Let

$$\mathcal{F}_{(0,0)} = \left\{ f : f(w;p) = \frac{\nabla_\theta F_\theta(S_{\{(0,0)\}|x;\theta})}{F_\theta(S_{\{(0,0)\}|x;\theta})} \right\},$$

$$\mathcal{F}_{(0,1)} = \left\{ f : f(w;p) = \frac{e_l' \nabla_\theta \eta_1(\theta;x)}{\eta_1(\theta;x)} \mathbb{I}_1(x;p) + \sum_{j=2}^{3} \frac{e_l'(\nabla_\theta \eta_1(\theta;x) - \nabla_\theta \eta_j(\theta;x))}{\eta_j(\theta;x)} \mathbb{I}_j(x;p), \right.$$
$$\left. l = 1, \ldots, d_\theta, , \ p \in \mathcal{H} \right\},$$

$$\mathcal{F}_{(1,0)} = \left\{ f : f(w;p) = \sum_{j=1}^{3} \frac{e_l' \nabla_\theta \eta_j(\theta;x)}{\eta_j(\theta;x)} \mathbb{I}_j(x;p), \ l = 1, \ldots, d_\theta, \ p \in \mathcal{H} \right\},$$

$$\mathcal{F}_{(1,1)} = \left\{ f : f(w;p) = \frac{\nabla_\theta F_\theta(S_{\{(1,1)\}|x;\theta})}{F_\theta(S_{\{(1,1)\}|x;\theta})} \right\}$$

where for each l, e_l denotes the l-th basis vector in $\mathbb{R}^{d_\theta}$.

Observe that the score satisfies $s_\theta(\cdot; p_{y|x}) \in \mathcal{F} \equiv \sum_{\bar{y} \in \mathcal{Y}} \mathcal{F}_{\bar{y}} \cdot 1\{y = \bar{y}\}$. In view of Theorem 2.10.6 (and Examples 2.10.7 and 2.10.10) in van der Vaart and Wellner (1996), to verify Assumption 3 (iv) it suffices to show that $\mathcal{F}_{\bar{y}}$ is P-Donsker for each $\bar{y}$. The P-Donskerness of $\mathcal{F}_{\bar{y}}$ for $\bar{y} = (0,0), (1,1)$ follows immediately from each set being a singleton and Assumption B.2 (b). Below, we show $\mathcal{F}_{(1,0)}$ is P-Donsker. The analysis for $\mathcal{F}_{(0,1)}$ is similar and is therefore omitted.

Discrete X: First, suppose that X is a vector of discrete random variables and Assumption B.2(a)(i) holds. For this setting, we show that $\mathcal{F}_{(1,0)}$ is a Vapnik-Chervonenkis (VC) class, which satisfies Pollard's uniform entropy condition.

Observe that $\mathcal{H}$ is finite-dimensional due to $\mathcal{Y} \times \mathcal{X}$ being finite. By Lemma 2.6.15 in van der Vaart and Wellner (1996), the VC-index of this class is $V(\mathcal{H}) \leq d_Y \times d_X + 2$, and hence $\mathcal{H}$ is a VC-class. The finite set of functions $\mathcal{E} = \{\eta_j(\theta, \cdot), \frac{\partial}{\partial \theta_k} \eta_j(\theta, \cdot), j = 1, \ldots, 3, k = 1, \ldots, d_\theta\}$ is also a VC-class. Note that $\mathcal{F}_{(1,0)}$ collects functions that can be

expressed as combinations of functions from $\mathcal{H}$ and $\mathcal{E}$ by multiplication, addition, division, and composition with an indicator function $1\{\cdot > 0\}$. By Lemma 2.6.18 in van der Vaart and Wellner (1996), $\mathcal{F}_{(1,0)}$ is a VC-class. Assumptions 1 and B.1(i) ensure that there is an envelope (constant) function $F = 3C/c$ such that $|f| \leq F$ for all $f \in \mathcal{F}_{(1,0)}$. By Theorem 2.5.2 in van der Vaart and Wellner (1996), $\mathcal{F}_{(1,0)}$ is a P-Donsker class.

Continuous X: Next, suppose that X is a vector of continuous random variables and Assumption B.2(a)(ii) holds. We show $\mathcal{F}_{(1,0)}$ is P-Donsker by verifying the conditions of Theorem 3 in Chen et al. (2003). For this, we first show the L^2-Hölder continuity of $f \in \mathcal{F}_{(1,0)}$ in p. In what follows, let

$$U_{l,j} = e_l' \nabla_\theta \eta_j(\theta; X)/\eta_j(\theta; X), \;\; l = 1, \ldots, d, \; j = 1, \ldots, 3.$$

By the triangle inequality,

$$\sup_{\|p-p'\|_{\mathcal{H}} \leq \delta} |f(w; p') - f(w; p)|^2 \leq \sup_{\|p-p'\|_{\mathcal{H}} \leq \delta} \sum_{j=1}^{3} u_{l,j}^2 |\mathbb{I}_j(x; p') - \mathbb{I}_j(x; p)|, \tag{B.28}$$

where $u_{l,j} = e_l' \nabla_\theta \eta_j(\theta; x)/\eta_j(\theta; x)$. Below, we focus on $u_{l,3}^2 |\mathbb{I}_3(x; p') - \mathbb{I}_3(x; p)|$, one of the terms in the sum on the right hand side of (B.28). The two other terms can be analyzed similarly. For δ sufficiently small,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\|p-p'\|_{\mathcal{H}} \leq \delta} U_{l,3}^2 |\mathbb{I}_3(X; p') - \mathbb{I}_3(X; p)| \Big] \\
&\qquad = \mathbb{E}\Big[\sup_{\|p-p'\|_{\mathcal{H}} \leq \delta} U_{l,3}^2 |1\{Z_2(X; p'_{y|x}) < 0\} - 1\{Z_2(X; p_{y|x}) < 0\}| \Big].
\end{aligned}$$

By Eq. (B.14) and $\sup_{x \in \mathcal{X}} |\eta_j(\theta; x)| \leq 1$, whenever $\|p'_{y|x} - p_{y|x}\|_{\mathcal{H}} \leq \delta$, we have

$$\sup_{x \in \mathcal{X}} |Z_j(x; p'_{y|x}) - Z_j(x; p_{y|x})| \leq 3\delta, \;\; j = 1, 2. \tag{B.29}$$

We next use the argument in Chen et al. (2003, p. 1600). Combining one side of Eq. (B.29), with the addition of a non-negative constant, we have $Z_2(x; p_{y|x}) - 3\delta \leq Z_2(x; p'_{y|x}) \leq$

$Z_2(x; p'_{y|x}) + 3\delta$, and hence

$$1\{Z_2(x; p_{y|x}) - 3\delta < 0\} \geq 1\{Z_2(x; p'_{y|x}) < 0\} \geq 1\{Z_2(x; p'_{y|x}) + 3\delta < 0\}. \tag{B.30}$$

Similarly, $Z_2(x; p'_{y|x}) - 3\delta \leq Z_2(x; p_{y|x}) \leq Z_2(x; p_{y|x}) + 3\delta$ implies

$$1\{Z_2(x; p'_{y|x}) - 3\delta < 0\} \geq 1\{Z_2(x; p_{y|x}) < 0\} \geq 1\{Z_2(x; p_{y|x}) + 3\delta < 0\}. \tag{B.31}$$

Combining (B.30)-(B.31), for any p', p with $\|p' - p\|_{\mathcal{H}} \leq \delta$,

$$\begin{aligned}|1\{Z_2(x; p'_{y|x}) < 0\} - 1\{Z_2(x; p_{y|x}) < 0\}| &\leq 1\{Z_2(x; p_{y|x}) - 3\delta < 0\} - 1\{Z_2(x; p_{y|x}) + 3\delta < 0\} \\ &\leq 1\{-3\delta < Z_2(x; p_{y|x}) < 3\delta\}\end{aligned}$$

where without loss of generality we assumed that $1\{Z_2(x; p_{y|x}) - 3\delta < 0\} - 1\{Z_2(x; p_{y|x}) + 3\delta < 0\} > 1\{Z_2(x; p'_{y|x}) - 3\delta < 0\} - 1\{Z_2(x; p'_{y|x}) + 3\delta < 0\}$. By the argument above, the law of iterated expectations, and Assumptions B.1(i), B.2(a)(ii), and B.2(b),

$$\begin{aligned}&\mathbb{E}\Big[\sup_{\|p-p'\|_{\mathcal{H}} \leq \delta} U_{l,3}^2 |1\{Z_2(X; p'_{y|x}) < 0\} - 1\{Z_2(X; p_{y|x}) < 0\}|\Big] \\ &\qquad \leq \mathbb{E}\Big[U_{l,3}^2 1\{-3\delta < Z_2(X; p_{y|x}) < 3\delta\}\Big] \\ &\qquad \leq \tfrac{C^2}{c^2} \int 1\{-3\delta < z_2 < 3\delta\} f_{Z_2}(z_2) dz_2 \leq K\delta,\end{aligned} \tag{B.32}$$

for some constant $K > 0$, where the last inequality follows from Assumption B.1 (iii). Applying a similar argument to the other two terms in (B.28), one can obtain

$$\mathbb{E}\Big[\sup_{\|p-p'\|_{\mathcal{H}} \leq \delta} |f(W; p') - f(W; p)|^2\Big]^{1/2} \leq K'\delta^{1/2}, \tag{B.33}$$

for some $K' > 0$. Hence f is L^2-Hölder continuous in p with Hölder exponent $1/2$.

Recall that $\mathcal{X}$ is a bounded convex subset of $\mathbb{R}^{d_X}$ with nonempty interior. By Theorem 2.7.1 in van der Vaart and Wellner (1996), $\ln N(\epsilon^2, \mathcal{C}_M^\alpha(\mathcal{X}), \|\cdot\|_\infty) \leq K\left(\frac{1}{\epsilon}\right)^{2d_X/\alpha}$ for some

$K > 0$. Note that $\mathcal{H} \subset (\mathcal{C}_M^\alpha(\mathcal{X}))^{\mathcal{Y}}$ and $|\mathcal{Y}| = 4$. For each $y \in \mathcal{Y}$, let $\{p_1(y|\cdot), \ldots, p_k(y|\cdot)\}$ be an ϵ^2-cover for $\mathcal{C}_M^\alpha(\mathcal{X})$ with respect to the sup norm. Then, $\{(p_{j_1}((0,0)|\cdot), p_{j_2}((0,1)|\cdot), p_{j_3}((1,0)|\cdot), p_{j_4}((1,1)|\cdot), j_l \in \{1,\ldots,k\}, l = 1,\ldots,4\}$ forms an ϵ^2-cover for $(\mathcal{C}_M^\alpha(\mathcal{X}))^{\mathcal{Y}}$ with respect to the maximum of the sup norms. Hence,

$$N(\epsilon^2, \mathcal{C}_M^\alpha(\mathcal{X})^{\mathcal{Y}}, \|\cdot\|_\infty) \leq e^{4K\left(\frac{1}{\epsilon}\right)^{2d_X/\alpha}}, \tag{B.34}$$

which in turn implies $\ln N(\epsilon^2, \mathcal{H}, \|\cdot\|_\infty) \leq 4K\left(\frac{1}{\epsilon}\right)^{2d_X/\alpha}$. Since $\alpha > d_X$, we have

$$\int_0^\infty \sqrt{\ln N(\epsilon^2, \mathcal{H}, \|\cdot\|_{\mathcal{H}})} d\epsilon < \infty. \tag{B.35}$$

We can now apply Theorem 3 in Chen et al. (2003), which ensures that $\mathcal{F}_{(1,0)}$ is P-Donsker.

Mixed X: Finally, suppose that X contains both continuous and discrete variables and Assumption B.2(a)(iii) holds. Again, we use Theorem 3 in Chen et al. (2003). We can argue as in the previous case, but (B.32) is modified as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{\|p-p'\|_{\mathcal{H}} \leq \delta} U_{l,3}^2 |1\{Z_2(X; p'_{y|x}) < 0\} - 1\{Z_2(X; p_{y|x}) < 0\}|\Big] \\
&\qquad \leq \mathbb{E}\Big[U_{l,3}^2 1\{-3\delta < Z_2(X; p_{y|x}) < 3\delta\}\Big] \\
&\qquad \leq \frac{C^2}{c^2}\mathbb{E}\left[\int 1\{-3\delta < z_2 < 3\delta\} f_{X_2|X_d}(z_2) dz_2\right] \leq K\delta,
\end{aligned} \tag{B.36}$$

for some constant $K > 0$, where the last inequality follows from Assumption B.1(iii). Therefore, (B.33) holds.

It remains to show (B.35). Recall that $N(\epsilon^2, \mathcal{C}_M^\alpha(\mathcal{X}), \|\cdot\|_\infty) \leq e^{K\left(\frac{1}{\epsilon}\right)^{2d_X/\alpha}}$ for some $K > 0$. Furthermore, $x_d \mapsto \ell_d(y_k|x_d)$ belongs to a finite-dmensional space $[-M, M]^{\mathcal{X}_d}$ with covering number satisfying $N(\epsilon^2, [-M, M]^{\mathcal{X}_d}, \|\cdot\|_\infty) \leq \left(\frac{\sqrt{2M}}{\epsilon}\right)^{2\dim(\mathcal{X}_d)}$. For each l, let $p_{c,1}(y_l|\cdot), \cdots, p_{c,N_1}(y_l|\cdot)$ be an ϵ^2-cover of $\mathcal{C}_M^\alpha(\mathcal{X}_c)$. Similarly, let $p_{d,1}(y_l|\cdot), \ldots, p_{d,N_2}(y_l|\cdot)$ be an ϵ^2-cover of $[-M, M]^{\mathcal{X}_d}$. Then, for any $p_{y|x} \in \mathcal{H}$ and $l \in \{1,\ldots,4\}$, there exist $k_1 \in \{1,\ldots,N_1\}$, $k_2 \in \{1,\ldots,N_2\}$, and $(\ell_c(y_k|\cdot), \ell_d(y_k|\cdot)) \in \mathcal{C}_M^\alpha(\mathcal{X}_c) \times [-M, M]^{\mathcal{X}_d}$ such

that

$$\begin{aligned}
&\sup_{x=(x_c',x_d')'\in\mathcal{X}_c\times\mathcal{X}_d}\left|p_{y_l|x}(y|x)-\phi_k(p_{c,k_1}(y_l|x_c),p_{d,k_1}(y_l|x_d))\right| \\
&\quad=\sup_{x=(x_c',x_d')'\in\mathcal{X}_c\times\mathcal{X}_d}\left|\phi_k(p_c(y_l|x_c),p_d(y_l|x_d))-\phi_k(p_{c,k_1}(y_l|x_c),p_{d,k_2}(y_l|x_d))\right| \\
&\qquad\le C\max\{\|p_c(y_l|\cdot)-p_{c,k_1}(y_l|\cdot)\|_\infty,\|p_d(y_l|\cdot)-p_{d,k_2}(y_l|\cdot)\|_\infty\}\le C\epsilon^2,
\end{aligned}$$

for some $0<C<\infty$ due to the Lipschitz continuity of ϕ_k. Therefore $\{(p_{c,k_1}(y_l|\cdot),p_{d,k_2}(y_l|\cdot))_{l=1}^4$, $k_1\in\{1,\dots,N_1\},k_2\in\{1,\dots,N_2\},l\in\{1,\dots,4\}\}$ is an $C\epsilon^2$-cover of $\mathcal{H}$. Hence,

$$N(\epsilon^2,\mathcal{H},\|\cdot\|_\infty)\le\left(N(\epsilon^2/C,\mathcal{C}_M^\alpha(\mathcal{X}),\|\cdot\|_\infty)\times N(\epsilon^2/C,[-M,M]^{\mathcal{X}_d},\|\cdot\|_\infty)\right)^4,$$

which in turn implies

$$\ln N(\epsilon^2,\mathcal{H},\|\cdot\|_\infty)\le 4K\left(\tfrac{\sqrt{C}}{\epsilon}\right)^{2d_X/\alpha}+8\dim(\mathcal{X}_d)\ln\left(\tfrac{\sqrt{2M}}{\epsilon}\right)\le K'\left(\tfrac{\sqrt{C}}{\epsilon}\right)^{2d_X/\alpha}$$

for some $K'>0$ for all ϵ small enough. Again, by $\alpha>d$, we obtain (B.35). This completes the proof of the proposition. *Q.E.D.*

We conclude this section by arguing that provided $\boldsymbol{X}$ has at least one component with continuous distribution, under Assumptions 3 (ii), B.1 (iii), and B.2 (b), the consistency of the covariance matrix estimator $\hat{\Sigma}_{n,\theta^*}$ required in Eq. (3.35) holds. This follows from Eq. (B.33), arguing as in Powell et al. (1989, Theorem 3.4), leveraging Assumption 3 (ii) and the fact that for $\bar{y}=(0,0),(1,1)$ the score does not depend on $p_{n,y|x}$ together with Assumption B.2 (b).

APPENDIX C: ADDITIONAL EXAMPLES

Example C.1 (Discrete choice with unobserved heterogeneity in choice sets)**.** Consider a discrete choice model, with a finite universe of alternatives $\mathcal{J}=\{1,\dots,J\}$. Let each alternative be characterized by a vector of covariates X_j, which might vary across decision mak-

ers, and let $X = [X_j, j \in \mathcal{J}]$. Let $U_j, j \in \mathcal{J}$, denote an unobserved characteristic of alternative j that varies across decision makers. As in the model proposed by Barseghyan et al. (2021), the decision maker draws a *choice set* $C \subseteq \mathcal{J}$ according to an unknown distribution, and chooses the alternative $Y \in C$ that maximizes their utility, denoted $\pi(X_j;\theta) + U_j$ for alternative j (for simplicity, assume ties occur with probability zero):[2]

$$Y = \arg\max_{j \in C} \left(\pi(X_j;\theta) + U_j\right).$$

The researcher observes (Y, X), but not C, and wishes to learn features of θ and the distribution of $(U_j, j \in \mathcal{J})$. Assume that $\mathbf{P}(|C| \geq \kappa) = 1$, for some known $\kappa \geq 2$; in words, this amounts to requiring that each decision maker draws a choice set of size at least κ, and that κ is known and larger than one.[3] For given $\theta \in \Theta$ and $x \in \mathcal{X}$, Barseghyan et al. (2021, Lemma A.1) show that the set of model implied optimal choices is a measurable correspondence (per Definition 1.1 in Molchanov and Molinari, 2018) given by the $J - \kappa + 1$ best alternatives in $\mathcal{J}$, so that

$$G(U|x;\theta) = \cup_{K \subseteq \mathcal{J}:|K|=\kappa} \left\{ \arg\max_{j \in K} \left(\pi(x_j;\theta) + U_j\right) \right\}.$$

We depict it in Panel (a) of Figure C.1, for $|\mathcal{J}| = 3$, as a function of $(u_1 - u_3, u_2 - u_3)$.

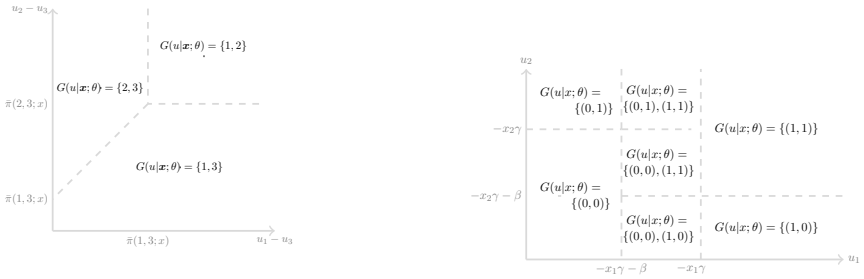
In this example, if $U_j, j \in \mathcal{J}$ has full support on $\mathbb{R}^J$, Assumption 1(b) is immediately satisfied with $\mathcal{A}_G = \{K \subset \mathcal{Y} : |K| \geq J - \kappa + 1\}$ because each set of alternatives in $\mathcal{J}$ of size $J - \kappa + 1$ can realize as the $J - \kappa + 1$ best. Assumption 1(c)-(d) can be verified similarly to how they are verified for Example 1.

It is also instructive to think about whether the introduction of a selection mechanism can allow for application of the method proposed in Chen et al. (2018) to this example.[4] Let $\mathbf{P}(Y_i = j|X_i;\theta,R)$ denote the model-implied conditional probability that alternative $j \in \mathcal{J}$

[2] As in Barseghyan et al. (2021), we can allow the utility function to be non-separable in (X, U), and we can allow U to include random coefficients that are decision maker specific and do not vary across alternatives.

[3] If $\mathbf{P}(|C| = 1) = 1$, the model has no empirical content and nothing can be learned about preferences.

[4] In the case of the entry game in Example 1, the selection mechanism in Eq. (3.3) can be integrated out against the distribution of U to obtain a function that plays the role of the nuisance parameter in Chen et al. (2018).



(a) Heterogeneous choice sets (b) Panel dynamic discrete choice

FIGURE C.1.—Stylized depictions of $G(\cdot|x;\theta)$ in our two examples. Notes: Panel (a) depicts Example C.1, with $\mathcal{J} = \{1,2,3\}$, $\kappa = 2$, and $\bar{\pi}(j,k;x) \equiv \pi(x_k;\theta) - \pi(x_j;\theta)$. Panel (b) depicts Example C.2 with $\beta \geq 0$.

is chosen given X_i and (θ, R), where $R(\cdot; X_i, U_i)$ denotes the conditional probability mass function of C_i given (X_i, U_i). For all $j \in \mathcal{J}$,

$$\mathbf{P}(Y_i = j|X_i;\theta,R) = \int \sum_{K \subseteq \mathcal{J}} \mathbf{1}\left(\arg\max_{k \in K} (\pi(X_k;\theta) + u_k) = j\right) R(K;X_i,u) dF_\theta.$$

To be able to apply Chen et al.'s (2018) method, one needs to further restrict the model and assume that R does not depend on U, in which case $R(\cdot; X_i)$ can come out of the integral. Doing so, however, severely restricts the class of models to which the procedure is applicable, since it requires the distribution of choice sets to be independent of preferences. Important examples of choice set formation mechanisms that violate this requirement include sequential search, rational inattention, and elimination by aspects (when the aspect with respect to which elimination occurs is the unobserved characteristic U_j). □

Example C.2 (Panel dynamic discrete choice)**.** Decision maker i chooses between actions $y = 0$ and $y = 1$ across multiple time periods, according to

$$Y_{it} = 1\{X_{it}\gamma + Y_{it-1}\beta + \alpha_i + \epsilon_{it} \geq 0\},\ i = 1,\ldots,n,\ t = 1,\ldots,T$$

with Y_{it} their decision in period t, X_{it} a vector of observed covariates in period t, α_i an individual-specific unobserved effect that is fixed over time, and ϵ_{it} an idiosyncratic

unobserved effect that varies over time. When $\beta \neq 0$, period t's choice depends on previous periods' choices, introducing state dependence. The researcher observes (Y_{it}, X_{it}) for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, but does not observe Y_{i0}, so that $\{Y_{i1}, \ldots, Y_{iT}\}$ is not fully determined and the model is incomplete (Heckman, 1978, Honoré and Tamer, 2006). Nonetheless, for given $(X_{it}, \alpha_i, \epsilon_{it}),\ t = 1, \ldots, T$, the model constrains the possible values that $(Y_{it}, t = 1, \ldots, T)$ can take. For example, with $T = 2$, one has:

$$Y_{i1} = \begin{cases} 1\{X_{i1}^\top \gamma + \alpha_i + \epsilon_{i1} \geq 0\} & \text{if } Y_{i0} = 0, \\ 1\{X_{i1}^\top \gamma + \beta + \alpha_i + \epsilon_{i1} \geq 0\} & \text{if } Y_{i0} = 1, \end{cases}$$

$$Y_{i2} = 1\{X_{i2}^\top \gamma + Y_{i1}\beta + \alpha_i + \epsilon_{i2} \geq 0\} \quad \text{if } Y_{i0} = 0 \text{ or } Y_{i0} = 1.$$

Denoting the unobservables as $U_{it} \equiv \alpha_i + \epsilon_{it}$, for given $\theta = (\gamma, \beta) \in \Theta$ and $x \in \mathcal{X}$, Chen and Kaido (2023) derive the correspondence $G(\cdot|x;\theta)$ as the set of values $(y_1, y_2) \in \{0,1\}^2$ that satisfy the above equations.[5] The correspondence is depicted in Panel (b) of Figure C.1 as a function of (u_1, u_2) for the case that $\beta \geq 0$. Similar examples arise in nonparametric models of state dependence (e.g., Torgovitsky, 2019). In this example, if the parameter space for β is a subset of $\mathbb{R}_{++}$, Assumption 1 (b) is satisfied because then for all $\theta \in \Theta$, $\mathcal{A}_G = \{\{(0,0)\}, \{(0,1)\}, \{(1,0)\}, \{(1,1)\}, \{(0,1),(1,1)\}, \{(0,0),(1,1)\}, \{(0,0),(1,0)\}\}$. Assumption 1 (c),(e) can be verified similarly to how they are verified for Example 1. □

APPENDIX: REFERENCES

BARSEGHYAN, LEVON, MAURA COUGHLIN, FRANCESCA MOLINARI, AND JOSHUA C. TEITELBAUM (2021): "Heterogeneous Choice Sets and Preferences," *Econometrica*, 89, 2015–2048. [22]

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. [24]

CHEN, XIAOHONG, TIMOTHY M. CHRISTENSEN, AND ELIE TAMER (2018): "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86, 1965–2018. [22, 23]

CHEN, XIAOHONG, OLIVER LINTON, AND INGRID VAN KEILEGOM (2003): "Estimation of Semiparametric Models When the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608. [18, 20]

[5] Using similar arguments as Molchanov and Molinari (2018, Example 1.5) it can be shown that this correspondence is measurable per Definition 1.1 in Molchanov and Molinari (2018).

GALICHON, ALFRED AND MARC HENRY (2011): "Set Identification in Models with Multiple Equilibria," *The Review of Economic Studies*, 78, 1264–1298. [4]

GAUVIN, JACQUES AND ROBERT JANIN (1990): "Directional derivative of the value function in parametric optimization," *Annals of Operations Research*, 27, 237–252. [8]

HECKMAN, JAMES J. (1978): "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence against the Hypothesis of Spurious State Dependence," *Annales de l'inséé*, 227–269. [24]

HONORÉ, BO E. AND ELIE TAMER (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74, 611–629. [24]

LUO, YE AND HAI WANG (2017): "Core Determining Class and Inequality Selection," *AER P&P*, 107, 274–77. [4]

MOLCHANOV, I. (2017): *Theory of Random Sets*, London: Springer, 2 ed. [1]

MOLCHANOV, ILYA AND FRANCESCA MOLINARI (2018): *Random Sets in Econometrics*, Econometric Society Monograph Series, Cambridge University Press, Cambridge UK. [1, 5, 22, 24]

PONOMAREV, KIRILL (2022): "Essays in Econometrics," Ph.D. thesis, UCLA. [4]

POWELL, JAMES L., JAMES H. STOCK, AND THOMAS M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430. [21]

ROCKAFELLAR, R. T. (1984): *Directional differentiability of the optimal value function in a nonlinear programming problem*, Berlin, Heidelberg: Springer Berlin Heidelberg, 213–226. [8]

TELGEN, JAN (1983): "Identifying Redundant Constraints and Implicit Equalities in Systems of Linear Constraints," *Management Science*, 29, 1209–1222. [4]

TORGOVITSKY, ALEXANDER (2019): "Nonparametric Inference on State Dependence in Unemployment," *Econometrica*, 87, 1475–1505. [24]

VAN DER VAART, A. W. AND JON A. WELLNER (1996): "Weak Convergence and Empirical Processes, With Applications to Statistics," *Springer Series in Statistics*. [11, 17, 18, 19]