Counterfactual Sensitivity and Robustness*

Timothy Christensen† Benjamin Connault‡

May 14, 2021

Abstract

We propose a framework for characterizing the sensitivity of counterfactuals with respect to parametric assumptions about the distribution of latent variables in a class of structural models. In particular, we show how to characterize the smallest and largest values of the counterfactual as the distribution of latent variables spans nonparametric neighborhoods of a researcher's parametric specification while other "structural" features of the model are maintained. Our procedure replaces the infinite-dimensional optimization with respect to the distribution by a finite-dimensional convex program and is therefore computationally simple to implement. We develop a novel MPEC implementation of our procedure to further simplify computation in models featuring endogenous parameters defined by equilibrium constraints. Our procedure recovers sharp bounds on the nonparametrically identified set of counterfactuals over large neighborhoods and has connections with local approaches to sensitivity analysis over small neighborhoods. We propose plug-in estimators of the smallest and largest counterfactuals and two procedures for inference. We illustrate the broad applicability of our procedure with empirical applications to matching models and dynamic discrete choice.

Keywords: Robustness, ambiguity, model uncertainty, misspecification, global sensitivity analysis.

JEL codes: C14, C18, C54, D81

*We are very grateful to R. Allen, S. Bonhomme, G. Chamberlain, A. Galvao, J. Lazarev, M. Masten, K. Menzel, D. Miller, M. Mogstad, F. Molinari, J. Nesbit, A. Poirier, J. Porter, T. Sargent, A. Torgovitsky, and numerous seminar and conference participants for helpful comments and suggestions. Maximilian Huber provided outstanding research assistance. Support from the National Science Foundation via grant SES-1919034 is gratefully acknowledged.

†Department of Economics, New York University. `timothy.christensen@nyu.edu`

‡IEX Group. `benjamin.connault@gmail.com`

1 Introduction

Researchers frequently make parametric assumptions about the distribution of latent variables (e.g. productivity shocks or random components of utility) when formulating structural models. Such assumptions are typically made for computational convenience[1] or because simulation-based methods are used for estimation. However, economic theory typically provides little guidance as to the correct specification of the distribution of latent variables and in many models, such as those we consider in this paper, the distribution is not nonparametrically identified. Ex ante policy experiments, or *counterfactuals*, can be particularly sensitive to such distributional assumptions. This sensitivity arises through two channels: the assumptions are used first at the estimation stage to define the model's mapping from structural parameters to observables (e.g. choice probabilities) and again when the model is solved under the policy intervention at the estimated structural parameters. The potential sensitivity of counterfactuals to such assumptions threatens the credibility of structural modeling exercises, a point made even by proponents of structural modeling (see, e.g., Keane, Todd, and Wolpin (2011)).

In this paper, we introduce a tractable econometric framework to characterize the sensitivity of counterfactuals with respect to parametric assumptions about the distribution of latent variables in a class of structural models. In particular, we show how to characterize the smallest and largest values of the counterfactual as the distribution of latent variables spans nonparametric neighborhoods of the researcher's assumed specification while other structural features of the model are maintained. This approach is in the spirit of global sensitivity analysis advocated by Leamer (1985). Global, rather than local, approaches to sensitivity analyses in nonlinear structural models are important, as policy interventions can have different effects at different points in the parameter space. Global sensitivity analyses of nonlinear models can be computationally and theoretically challenging, however. Local sensitivity analyses—based on linearizing small perturbations around a correct specification—are often more tractable. However, local approaches may fail to correctly characterize the counterfactuals predicted by the model when the researcher's parametric assumption is misspecified by a degree that is not vanishingly small. This is particularly important for the class of problems we consider, in which the distribution of latent variables is not nonparametrically identified.

The central innovation of our procedure is to borrow from the robustness literature in economics pioneered by Hansen and Sargent (2001, 2008) to simplify computation using convex programming. Following the robustness literature, we define nonparametric neighborhoods in

[1]For instance, work on static discrete choice following McFadden (1974) often assumes random components of utility are Gumbel-distributed as this leads to closed-form expressions for choice probabilities and welfare measures. Similarly, dynamic discrete choice models following Rust (1987) are typically implemented assuming that latent payoff shocks are Gumbel-distributed for computational convenience. Additionally, models of static or dynamic discrete games often impose parametric assumptions about the distribution of payoff shocks (see, e.g., Aguirregabiria and Mira (2007), Bajari, Benkard, and Levin (2007), Ciliberto and Tamer (2009)).

terms of statistical divergence from the researcher's assumed specification, with the option to add location or scale normalizations and certain shape restrictions. For tractability, we restrict attention to a class of models that may be written as a finite number of moment in/equalities, where the expectation is taken with respect to the distribution of unobservables. Though somewhat restrictive, this class is sufficiently broad that it accommodates many models of static and dynamic discrete choice, static or dynamic discrete games, and matching markets.

To briefly summarize our procedure, consider the problem of minimizing or maximizing the counterfactual at a particular value of structural parameters by varying the distribution over the neighborhood, subject to the model's in/equality restrictions. This *inner* optimization problem is infinite-dimensional, but can be recast as convex program of fixed (finite) dimension. The value of the convex program is treated as a criterion function, which is optimized in an *outer* optimization with respect to structural parameters.

Importantly, our procedure recovers sharp bounds on the nonparametrically identified set of counterfactuals as the neighborhood size becomes large. Neighborhoods constrained by statistical divergence can therefore be viewed as a type of (infinite-dimensional) sieve: although they exclude many distributions, as their size increases the neighborhoods span all distributions relevant for characterizing the nonparametrically identified set of counterfactuals. Unlike sieve methods based on parametric families of growing dimension, here the dimension of the optimization problem remains fixed as we consider increasingly rich classes of distributions. Our methods therefore provide a tractable way for characterizing nonparametrically identified sets of counterfactuals in nonlinear structural models.

In addition, we develop a novel MPEC implementation of our procedure to further simplify computation in models featuring endogenous parameters defined by equilibrium constraints (e.g. value functions in dynamic discrete choice models). The novelty relative to other MPEC implementations (e.g. Su and Judd (2012)) is that the equilibrium constraints are evaluated under a "least favorable" distribution arising from the solution to the inner optimization. We show that this implementation can produce significant computational gains for dynamic discrete choice models.

For estimation and inference, we propose plug-in estimators for the lower and upper bounds on the counterfactual over nonparametric neighborhoods. We show the estimators are consistent and establish their (nonstandard) asymptotic distribution. In addition, we propose two procedures for inference: a computationally inexpensive projection-based procedure and a more efficient bootstrap-based procedure. Our methodology, estimators, and inference procedures are robust to partial identification and irregular estimability of structural parameters, both of which may be important in applications.

We illustrate our procedure with empirical applications to matching models with transferable utility (Choo and Siow, 2006) in which we revisit Chiappori, Salanié, and Weiss (2017) and to welfare analysis in models of dynamic discrete choice (Rust, 1987). Both applications feature several hundred moments, illustrating the computational feasibility of our procedure.

Related literature. Our approach has connections with global prior sensitivity in Bayesian analysis (Chamberlain and Leamer, 1976; Leamer, 1982; Berger, 1984; Berger and Berliner, 1986). In particular, Giacomini, Kitagawa, and Uhlig (2016) and Ho (2018) consider sets of priors constrained by Kullback–Leibler divergence relative to a default prior.

Motivated by questions of sensitivity, Chen, Tamer, and Torgovitsky (2011) study inference in partially identified semiparametric likelihood models using sieve approximations for the infinite-dimensional parameter (the distribution of latent variables in our context). Sieve methods would require a (typically non-convex) optimization over the sieve coefficients, whose number must increase to infinity to recover the nonparametrically identified set. For the class of moment-based problems we consider, our approach eliminates the infinite-dimensional nuisance parameter via a convex program of fixed dimension.

Several other works have used convex duality to characterize identification regions for models with latent variables. Most closely related are Ekeland, Galichon, and Henry (2010) and Schennach (2014).[2] The primal problem we study involves minimizing or maximizing a counterfactual with respect to a distribution in a neighborhood of a researcher's parametric specification subject to moment in/equalities. This is a different problem from that which is studied in these works and, consequently, its dual formulation is different. Moreover, our focus is on counterfactuals and our estimation and inference methods are tailored accordingly.

Torgovitsky (2019b) uses linear programming to characterize sharp identified sets in a class of latent variable models defined by finitely many quantile restrictions. His approach may be more computationally convenient than ours within this class of problems as it uses linear programming. However, several important moment functions or counterfactuals cannot be expressed as quantile restrictions, including measures of social surplus (or welfare) in discrete choice following McFadden (1978) and Bellman equations in dynamic discrete choice models. Our approach is compatible with these, and therefore allows for the construction of identified sets in broader classes of model, as well as addressing issues of sensitivity.

There is also work on nonparametrically identified sets of counterfactuals in some specific models with latent variables. Examples include Manski (2007, 2014), Allen and Rehbeck (2019), Chiong, Hsieh, and Shum (2017), Tebaldi, Torgovitsky, and Yang (2019), Lafférs (2019) and Torgovitsky (2019a). Most closely related to our work, Norets and Tang (2014) construct identified sets of counterfactual conditional choice probabilities (CCPs) in dynamic binary choice models without parametric assumptions on latent payoff shocks. Their approach, which

[2]Other works using various "duality" notions to simplify the construction of identified sets in general classes of models with latent variables include Beresteanu, Molchanov, and Molinari (2011), Galichon and Henry (2011), Chesher and Rosen (2017), and Torgovitsky (2019b). In recent work that is concurrent with ours, Li (2018) relaxes some restrictions on the moment functions and the support of unobservables in Ekeland et al. (2010).

uses linear programming, is specific to counterfactual CCPs in dynamic binary choice models while we allow for more general counterfactuals (e.g. welfare) and multinomial choice. We also accommodate a broader range of normalizations and shape restrictions on the distribution, as well as addressing issues of sensitivity.[3]

Finally, a recent literature on local sensitivity considers deviations from a well-specified model that shrink to zero at an appropriate rate with the sample size; see, e.g., Kitamura, Otsu, and Evdokimov (2013), Andrews, Gentzkow, and Shapiro (2017, 2020), Armstrong and Kolesár (2021), Bonhomme and Weidner (2018), and Mukhin (2018). The typical justification for vanishingly small neighborhoods is that larger departures may be detected using various specification tests as one observes more data. This motivation is not compelling in our setting as the distribution of unobservables is not nonparametrically identified. Moreover, much of this recent literature is concerned with local misspecification of the moment conditions, which is a different problem from that which we consider.

The remainder of the paper is organized as follows. Section 2 describes our procedure, introduces the estimators, and shows that our procedure can be used to obtain sharp bounds on the nonparametrically identified set of counterfactuals. Section 3 discusses practical aspects, including computation and our MPEC reformulation. Section 4 presents theoretical results and practical guidance to interpret the neighborhood size. Empirical applications are presented Section 5. Section 6 discusses estimation and inference and Section 7 discusses connections with local approaches to sensitivity analysis. The appendix presents extensions of our methodology, supplemental results, and all proofs.

2 Procedure

2.1 Setup

We consider a class of models that links a structural parameter $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$, a vector of targeted moments $P_0 \in \mathcal{P} \subseteq \mathbb{R}^{d_P}$, and possibly an auxiliary parameter $\gamma_0 \in \Gamma$ (a metric space) via the moment restrictions

$$\mathbb{E}^F[g_1(U, \theta, \gamma_0)] \leq P_{10}, \tag{1a}$$

$$\mathbb{E}^F[g_2(U, \theta, \gamma_0)] = P_{20}, \tag{1b}$$

$$\mathbb{E}^F[g_3(U, \theta, \gamma_0)] \leq 0, \tag{1c}$$

$$\mathbb{E}^F[g_4(U, \theta, \gamma_0)] = 0, \tag{1d}$$

[3] Kalouptsidi, Scott, and Souza-Rodrigues (2017) and Kalouptsidi, Kitamura, Lima, and Souza-Rodrigues (2020) consider the converse problem, in which flow payoffs are nonparametric (as they can be in our setting) but the distribution of latent payoff shocks is known. Note, however, that this distribution is not nonparametrically identified when the state-space is discrete.

where $g_1, \ldots, g_4$ are vectors of moment functions, $P_0 = (P_{10}, P_{20})$ is partitioned conformably, U is a vector of latent variables, and the expectations are taken with respect to the distribution F of U. We assume that the researcher has consistent estimators $(\hat{P}, \hat{\gamma})$ of (P_0, γ_0). We also assume that the researcher is interested in a (scalar) counterfactual of the form

$$\kappa = \mathbb{E}^F[k(U, \theta, \gamma_0)] \,. \tag{2}$$

Several workhorse models and counterfactuals of interest are subsumed in this framework.

Example 2.1 (Discrete choice and consumer welfare) Suppose an individual i derives utility $h_j(X_{ij}, \theta) + U_{ij}$ from consuming good $j \in \mathcal{J}_0$, where $\mathcal{J}_0 = \{0\} \cup \mathcal{J}$ with $\mathcal{J} = \{1, \ldots, J\}$, $X_i = (X_{ij})_{j \in \mathcal{J}_0}$ is a vector of observable characteristics, and the latent random vector $(U_{ij})_{j \in \mathcal{J}_0}$ represents the components of individual i's utilities that are unobserved by the econometrician.[4] Following McFadden (1974), it is common to assume that $(U_{ij})_{j \in \mathcal{J}_0}$ are independent of X_i and i.i.d. across individuals i with continuous distribution F. The (conditional) probability that an individual i for whom $X_i = x$ chooses j is then

$$p(j|x) = \mathbb{P}_F\left(h_j(x, \theta) + U_j \geq \max\nolimits_{j' \in \mathcal{J}_0}\left(h_{j'}(x, \theta) + U_{j'}\right)\right), \quad j \in \mathcal{J}_0 \,, \tag{3}$$

where $\mathbb{P}_F$ denotes probabilities under the distribution F of $U := (U_j)_{j \in \mathcal{J}_0}$. In empirical work, θ is typically estimated from choice data by matching empirical choice probabilities with model-implied choice probabilities under a parametric functional form for F (e.g. i.i.d. Gumbel, which leads to multinomial logit expression for choice probabilities).

Welfare analyses are often based on the social surplus (McFadden, 1978)

$$W(x) := \mathbb{E}^F\left[\max_{j \in \mathcal{J}_0}\left(h_j(x, \theta) + U_j\right)\right],$$

which represents the average utility that consumers with characteristics $X = x$ derive from their choice problem, or the change in surplus $\Delta W(x_a, x_b) = W(x_a) - W(x_b)$ associated with a shift from x^b to x^a. As $W(x)$ is not identified from aggregate choice data, welfare analyses can be sensitive to researchers' assumptions about F.[5]

Our approach may be used to examine sensitivity of $W(x)$ and $\Delta W(x_a, x_b)$ to assumptions about F in models in which the support $\mathcal{X}$ of X is a finite set. A leading example is empirical matching models in which $\mathcal{X}$ indexes agents' types (see, e.g., Dagsvik (2000), Choo and Siow (2006), Chiappori et al. (2017), and Section 5.1). In our notation, g_2 is a vector of indicator

[4]Our approach applies equally in settings in which variables are indexed by individual, choice, and market. We maintain this simpler presentation (indexing by individual and choice only) to simplify notation.

[5]Allen and Rehbeck (2019) provide sufficient conditions for nonparametric identification of $\Delta W(x_a, x_b)$ from aggregate choice data when x_a and x_b lie in the support of X and the distribution of X satisfies some continuity conditions. Note that $\Delta W(x_a, x_b)$ is not nonparametrically identified when all characteristics are discrete.

functions representing the choice probabilities (3):

$$g_2(U,\theta) = \left(1\!\!1 \left\{ h_j(x,\theta) + U_j \geq \max_{j' \in \mathcal{J}_0} \left(h_{j'}(x,\theta) + U_{j'} \right) \right\} \right)_{(j,x) \in \mathcal{J} \times \mathcal{X}}$$

and $P_{20} = (\Pr(j|x))_{(j,x) \in \mathcal{J} \times \mathcal{X}}$ is the vector of true choice probabilities ($j = 0$ is redundant). There is no g_1, g_3, or γ in this model. The welfare measures $W(x)$ and $\Delta W(x_a, x_b)$ are of the form (2) with $k(U,\theta) = \max_{j \in \mathcal{J}_0} (h_j(x,\theta) + U_j)$ and $k(U,\theta) = \max_{j \in \mathcal{J}_0} (h_j(x_a,\theta) + U_j) - \max_{j \in \mathcal{J}_0} (h_j(x_b,\theta) + U_j)$, respectively. □

Example 2.2 (Discrete games) Consider a complete-information, two-player game following Bresnahan and Reiss (1990, 1991), Berry (1992), and Tamer (2003):

		Firm 2	
		0	1
Firm 1	0	$(0,0)$	$(0, \beta_2' x + U_2)$
	1	$(\beta_1' x + U_1, 0)$	$(\beta_1' x - \Delta_1 + U_1, \beta_2' x - \Delta_2 + U_2)$

Table 1: Payoff matrix for (Firm 1, Firm 2) when $X = x$.

We focus on static, two-player, complete-information games to simplify exposition, but our procedure can be applied more broadly. In Table 1, $U = (U_1, U_2)$ is the latent (to the econometrician) component of firms' profits which is typically assumed to be independent of covariates X. Suppose that the solution concept is restricted to equilibria in pure strategies. The econometrician may estimate the probabilities of the potential market structures $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$ (conditional on X) form data on a large number of markets. Parameters are estimated by matching the observed and model-implied probabilities. As the model is incomplete—there are certain realizations of U for which there are two equilibria in pure strategies (Firm 1 enters and Firm 2 does not, or vice versa)—moment inequality methods are typically used to be robust to potential misspecification of the equilibrium selection mechanism. However, strong parametric assumptions are often made about the distribution of U (typically bivariate Normal) to derive model-implied probabilities; see, e.g., Berry (1992), Ciliberto and Tamer (2009), Beresteanu et al. (2011), and Kline and Tamer (2016). Given the emphasis on robustness with respect to equilibrium selection, it seems natural to also question the sensitivity of counterfactuals to parametric assumptions for U.

This model falls into our setup when the regressors X have finite support $\mathcal{X}$.[6] In our notation, g_1 collects the moment inequalities that bound the probabilities of $(0,1)$ and $(1,0)$

[6] Continuous regressors are often discretized in empirical applications; see Ciliberto and Tamer (2009), Grieco (2014), Kline and Tamer (2016).

with P_{10} denoting the corresponding true probabilities:

$$g_1(U,\theta) = \begin{bmatrix} (-1\!\!1\{U_1 \geq -\beta_1' x; U_2 \leq \Delta_2 - \beta_2' x\})_{x\in\mathcal{X}} \\ (-1\!\!1\{U_1 \leq \Delta_1 - \beta_1' x; U_2 \geq -\beta_2' x\})_{x\in\mathcal{X}} \end{bmatrix}, \quad P_{10} = \begin{bmatrix} (-\Pr((1,0)|X=x))_{x\in\mathcal{X}} \\ (-\Pr((0,1)|X=x))_{x\in\mathcal{X}} \end{bmatrix}.$$

Similarly, g_2 and P_{20} collect the moment conditions and probabilities for $(0,0)$ and $(1,1)$:

$$g_2(U,\theta) = \begin{bmatrix} (1\!\!1\{U_1 \leq -\beta_1' x;\ U_2 \leq -\beta_2' x\})_{x\in\mathcal{X}} \\ (1\!\!1\{U_1 \geq \Delta_1 - \beta_1' x;\ U_2 \geq \Delta_2 - \beta_2' x\})_{x\in\mathcal{X}} \end{bmatrix}, \quad P_{20} = \begin{bmatrix} (\Pr((0,0)|X=x))_{x\in\mathcal{X}} \\ (\Pr((1,1)|X=x))_{x\in\mathcal{X}} \end{bmatrix}.$$

There is no g_3 or γ in this model. The vector of structural parameters is $\theta = (\Delta_1, \Delta_2, \beta_1, \beta_2)$.[7] In terms of counterfactuals, Ciliberto and Tamer (2009) compute upper bounds on the entry probability of entrants under a counterfactual payoff shift, say $\tau(\theta)$. In the present context, the k function corresponding to the upper bound on the probability of observing firm 1 in a market with $X = x$ is $k(U,\theta) = 1\!\!1\{U_1 \geq \tau(\theta) - \beta_1' x\}$. Average maximum entry probabilities across markets can be computed similarly. □

Example 2.3 (Dynamic discrete choice) Consider a canonical dynamic discrete choice (DDC) model following Rust (1987). The decision maker solves

$$V(s) = \mathbb{E}^F \left[\max_{d\in\mathcal{D}_0} \left(\pi_{d,s}(\theta_\pi) + U_d + \beta E[V(s')|s,d] \right) \right], \tag{4}$$

where s is an observed Markov state variable with finite state-space $\mathcal{S}$, $\mathcal{D}_0 = \{0\} \cup \mathcal{D}$ with $\mathcal{D} = \{1,\ldots,D\}$ is the set of actions, $\pi_{d,s}$ is the flow payoff for action d in state s which is indexed by parameters θ_π, U_d is a utility shock observed by the agent but not the econometrician, and $E[\,\cdot\,|s,d]$ denotes expectation with respect to the future state. The vector $U = (U_0, U_1, \ldots, U_D)$ is continuously distributed independently of the state with distribution F. The CCP of action d in state s is

$$p(d|s) = \mathbb{P}_F \left(\pi_{d,s}(\theta_\pi) + U_d + \beta E[V(s')|s,d] \geq \max_{d'\in\mathcal{D}_0} \left(\pi_{d',s}(\theta_\pi) + U_{d'} + \beta E[V(s')|s,d'] \right) \right), \tag{5}$$

where $\mathbb{P}_F$ denotes probabilities under the distribution F of U.

In empirical work, CCPs and transition distributions for s are estimated from data on (s,d). The parameters θ_π or (θ_π, β) are then estimated under a parametric assumption on F to fit the model-implied CCPs to the observed choice data. It is common to assume the payoff shocks are i.i.d. (across s and d) standard Gumbel, which yields closed-form expressions for the right-hand sides of (4) and (5). Counterfactuals are computed by first solving (4) under

[7] We can also allow the correlation between U_1 and U_2 under a bivariate normal parametric specification to be treated as a free parameter. Let U_1 and U_2 be independent standard normal, replace U_1 and U_2 in the above displays with V_1 and V_2 where $V_1 = U_1$, $V_2 = \rho U_1 + \sqrt{1-\rho^2} U_2$, and let $\theta = (\Delta_1, \Delta_2, \beta_1, \beta_2, \rho)$.

alternative laws of motion, flow payoffs, or sets of actions.

Our procedure may be applied to examine sensitivity of counterfactuals to researchers' assumptions about F as follows. Let $\theta = (\theta_\pi, \beta, v, \tilde{v})$, where $v = (V(s))_{s\in\mathcal{S}}$ and $\tilde{v} = (\tilde{V}(s))_{s\in\mathcal{S}}$ are the pre- and post-intervention value functions. In our notation, $\gamma = (M_d)_{d\in\mathcal{D}_0}$ collects the transition matrices for s, g_2 collects indicator functions representing the CCPs (5):

$$g_2(U,\theta,\gamma) = \left(1\!\!1\left\{\pi_{d,s}(\theta_\pi) + U_d + \beta M_d v \geq \max_{d'\in\mathcal{D}_0}\left(\pi_{d',s}(\theta_\pi) + U_{d'} + \beta M_{d'} v\right)\right\}\right)_{(d,s)\in\mathcal{D}\times\mathcal{S}},$$

$P_{20} = (\Pr(d|s))_{(d,s)\in\mathcal{D}\times\mathcal{S}}$ is the corresponding vector of true CCPs ($d = 0$ is redundant), and g_4 is a vector of moment functions representing (4) pre- and post-intervention:

$$g_4(U,\theta,\gamma) = \left[\begin{array}{c} \left(\max_{d\in\mathcal{D}_0}\left\{\pi_{d,s}(\theta_\pi) + U_d + \beta M_d v\right\} - v_s\right)_{s\in\mathcal{S}} \\ \left(\max_{d\in\tilde{\mathcal{D}}_0}\left\{\tilde{\pi}_{d,s}(\theta_\pi) + U_d + \tilde{\beta}\tilde{M}_d \tilde{v}\right\} - \tilde{v}_s\right)_{s\in\mathcal{S}} \end{array}\right], \tag{6}$$

where $v_s = V(s)$, $\tilde{v}_s = \tilde{V}(s)$, and $\tilde{\mathcal{D}}_0$, $\tilde{u}$, $\tilde{\beta}$, $\tilde{M}_d$ denote the counterfactual action set, flow payoffs, discount factor, and law of motion. There are several possibilities for k from (2). For the counterfactual CCP for action d in state s one could use

$$k(U,\theta,\gamma) = 1\!\!1\left\{\tilde{\pi}_{d,s}(\theta_\pi) + U_d + \tilde{\beta}\tilde{M}_d\tilde{v} \geq \max_{d'\in\tilde{\mathcal{D}}_0}\left(\tilde{\pi}_{d',s}(\theta_\pi) + U_{d'} + \tilde{\beta}\tilde{M}_{d'}\tilde{v}\right)\right\},$$

whereas for change in average welfare one could use $k(u,\theta,\gamma) = k(\theta,\gamma) = w'(\tilde{v} - v)$ for a weight vector w. □

Remark 2.1 Our setup requires that the moments do not depend on the data beyond their dependence through γ and P. Our setup permits conditional moments models of the form $\mathbb{E}[g_1(U,X,\theta,\gamma)|X = x] \leq P_{10}(x)$ (and similarly for the other moment conditions involving g_2, g_3, and g_4) provided U is independent of X and X takes values in a finite set, in which case the moment functions at each value of X are stacked to form g_1, g_2, g_3, g_4 (see Examples 2.1-2.3).

Remark 2.2 Appendix A presents extensions of our setup to (i) conditional moment restriction models where the distribution of $U|X$ may vary with X provided X takes values in a finite set, and (ii) nonseparable moments of the form $\mathbb{E}[g_1(X,U,\theta,\gamma)] \leq P_{10}$ (and similarly for the other moment conditions involving g_2, g_3, and g_4) with discrete X. Although models with continuous covariates fall outside the scope of our procedure, continuous covariates can, in principle, be discretized up to the limits of one's computing resources.

Remark 2.3 Our setup relies on the counterfactual being scalar and of the form (2). If k is vector-valued then our procedure can be applied to compute the support function of the identified set of counterfactuals:[8] for any conformable unit vector τ set $k^\tau(U,\theta,\gamma) = \tau' k(U,\theta,\gamma)$ and replace (2) with $\kappa^\tau = \mathbb{E}^F[k^\tau(U,\theta,\gamma_0)]$. Our setup excludes counterfactuals that are infinite-dimensional, however, such as the distribution of the number of firms in a market.

The distribution F is not nonparametrically identified in any of the above examples or, more generally, in the class of problems we consider. Therefore, in common practice, a seemingly reasonable or computationally convenient distribution, say F_*, is assumed by the researcher and maintained throughout the analysis (e.g. bivariate Normal in Example 2.2 and i.i.d. Gumbel in Examples 2.1 and 2.3). Given F_* and estimates $\hat{P} = (\hat{P}_1, \hat{P}_2)$ of P_0 and possibly $\hat{\gamma}$ of γ_0, the researcher would estimate θ using a criterion function based on the moment conditions

$$\begin{aligned} &\mathbb{E}^{F_*}[g_1(U,\theta,\hat{\gamma})] \le \hat{P}_1\,, \quad \mathbb{E}^{F_*}[g_2(U,\theta,\hat{\gamma})] = \hat{P}_2\,, \\ &\mathbb{E}^{F_*}[g_3(U,\theta,\hat{\gamma})] \le 0\,, \quad\;\; \mathbb{E}^{F_*}[g_4(U,\theta,\hat{\gamma})] = 0\,. \end{aligned} \tag{7}$$

Given any such estimator $\hat{\theta}$, the counterfactual κ can then be estimated using

$$\hat{\kappa} = \mathbb{E}^{F_*}[k(U,\hat{\theta},\hat{\gamma})]\,.$$

If the function k does not depend on U, then we simply have $\kappa = k(\theta,\gamma_0)$ and $\hat{\kappa} = k(\hat{\theta},\hat{\gamma})$. In this case the estimated counterfactual $\hat{\kappa}$ will still depend *implicitly* on F_* through $\hat{\theta}$. While the preceding discussion has assumed point identification of θ and κ for sake of exposition, our methods allow structural parameters and counterfactuals to be partially identified.

In the preceding description of a structural modeling exercise, the researcher's parametric specification F_* is being used both for estimation of the structural parameter θ and again for computation of the counterfactual κ. A natural question that arises is: to what extent does the counterfactual depend on the researcher's choice F_*, and to what extent does it depend on the underlying structure of the model? The main contribution of this paper is to provide a tractable econometric framework to address this question.

2.2 Our Approach

In the spirit of sensitivity analysis, we relax the parametric assumption F_* and allow F to vary over nonparametric neighborhoods $\mathcal{N}_\delta$ of F_*, where δ is a measure of "size". When we do so, there are pairs $(\theta, F) \in \Theta \times \mathcal{N}_\delta$ that satisfy the moment conditions (1), but which may yield different values of the counterfactual. We wish to characterize the smallest and largest values

[8]The support function $s_A : \mathbb{R}^k \to \mathbb{R}$ of $A \subset \mathbb{R}^k$ is $s_A(x) = \sup_{a \in A} a'x$. Note that $s_A = s_B$ if and only if the closures of the convex hulls of A and B are equal.

of the counterfactual over such (θ, F) pairs:

$$\underline{\kappa}_\delta = \inf_{\theta \in \Theta, F \in \mathcal{N}_\delta} \mathbb{E}^F[k(U, \theta, \gamma_0)] \quad \text{subject to (1)}, \tag{8}$$

$$\overline{\kappa}_\delta = \sup_{\theta \in \Theta, F \in \mathcal{N}_\delta} \mathbb{E}^F[k(U, \theta, \gamma_0)] \quad \text{subject to (1)}. \tag{9}$$

By focusing on $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$, our approach does not require point-identification of θ under F_* or any other candidate distribution F. Thus, we accommodate models with partially-identified structural parameters and counterfactuals. Our approach also sidesteps having to compute the identified set of structural parameters.

Computing $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ requires solving infinite-dimensional optimization problems with respect to F. This is made tractable for the class of models we consider by a convenient choice of $\mathcal{N}_\delta$. Following work on ambiguity and model uncertainty by Hansen and Sargent (2001) and Maccheroni, Marinacci, and Rustichini (2006), we consider neighborhoods that are constrained by ϕ-divergence from F_*:

$$\mathcal{N}_\delta = \{F \in \mathcal{F} : D_\phi(F \| F_*) \leq \delta\}, \tag{10}$$

with

$$D_\phi(F \| F_*) = \begin{cases} \displaystyle\int \phi\left(\frac{\mathrm{d}F}{\mathrm{d}F_*}\right) \mathrm{d}F_* & \text{if } F \ll F_*, \\ +\infty & \text{otherwise}, \end{cases}$$

where $\mathcal{F}$ denotes all probability measures on $\mathcal{U}$ (the support of U) and $F \ll F_*$ denotes absolute continuity of F relative to F_*. The function $\phi : [0, \infty) \to \mathbb{R} \cup \{+\infty\}$ is a non-negative convex function that penalizes deviations of F from F_*. For example, $\phi(x) = x \log x - x + 1$ corresponds to Kullback–Leibler (KL) divergence, $\phi(x) = \frac{1}{2}(x-1)^2$ corresponds to Pearson χ^2 divergence, and

$$\phi(x) = \frac{x^p - 1 - p(x-1)}{p(p-1)}, \quad (p > 1),$$

corresponds to L^p divergence. Common choices for F_* have positive (Lebesgue) density, so the absolute continuity condition merely rules out F with mass points.

The neighborhoods may be further disciplined by incorporating normalizations, smoothness constraints, or other shape restrictions in $g_1, \ldots, g_4$. Examples include: (i) location normalizations, e.g. $\mathbb{E}^F[U] = 0$ or $\mathbb{E}^F[1\!\!1\{U_i \leq 0\} - 0.5] = 0$; (ii) scale normalizations, e.g. $\mathbb{E}^F[U_i^2] = 1$; (iii) covariance normalizations, e.g. $\mathbb{E}^F[UU'] = I$, or bounds, e.g. $\mathbb{E}^F[UU'] \leq \Sigma$; (iv) smoothness restrictions, e.g. $\mathbb{E}^F[1\!\!1\{U_i \leq a_{k+1}\} - 1\!\!1\{U_i \leq a_k\}] \leq C$ for $a_1 < \ldots < a_K$ and a positive constant C, for each element U_i of U; and (v) notions of exchangeability. Exchangeability can carry important economic content in certain applications and is easy to impose whenever F_* is exchangeable (see Appendix A.3). Researchers may add or remove such restrictions to investigate how these affect the sets of counterfactuals.

2.3 Dual Formulation

In this subsection we use convex duality to simplify computation of $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$. We start by noting $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ may be written as the solution to two profiled optimization problems:

$$\underline{\kappa}_\delta = \inf_{\theta \in \Theta} \underline{K}_\delta(\theta; \gamma_0, P_0)\,, \qquad\qquad \overline{\kappa}_\delta = \sup_{\theta \in \Theta} \overline{K}_\delta(\theta; \gamma_0, P_0)\,,$$

where the criterion functions $\underline{K}_\delta(\theta; \gamma_0, P_0)$ and $\overline{K}_\delta(\theta; \gamma_0, P_0)$ are, respectively, the infimum and supremum of $\mathbb{E}^F[k(U, \theta, \gamma_0)]$ with respect to $F \in \mathcal{N}_\delta$ subject to the moment conditions (1). In what follows, it is helpful to define the criterion functions at a generic (γ, P). To do so, we say that the moment conditions (1) hold "at (θ, γ, P)" if they hold when γ_0 is replaced by γ and P_0 is replaced by P. Then

$$\underline{K}_\delta(\theta; \gamma, P) = \inf_{F \in \mathcal{N}_\delta} \mathbb{E}^F[k(U, \theta, \gamma)] \quad \text{subject to (1) holding at } (\theta, \gamma, P)\,, \tag{11}$$

$$\overline{K}_\delta(\theta; \gamma, P) = \sup_{F \in \mathcal{N}_\delta} \mathbb{E}^F[k(U, \theta, \gamma)] \quad \text{subject to (1) holding at } (\theta, \gamma, P)\,, \tag{12}$$

with the understanding that $\underline{K}_\delta(\theta; \gamma, P) = +\infty$ and $\overline{K}_\delta(\theta; \gamma, P) = -\infty$ if there does not exist a distribution in $\mathcal{N}_\delta$ for which the moment conditions (1) hold at (θ, γ, P).

To justify the dual formulation, we first impose some mild regularity conditions on F_*, ϕ, and the moment functions. The conditions are similar to those justifying duality results in generalized empirical likelihood estimation (see, e.g., Komunjer and Ragusa (2016)).

Definition 2.1 Φ_0 *consists of all* $\phi : [0, \infty) \to \mathbb{R} \cup \{+\infty\}$ *such that* ϕ *is twice continuously differentiable on* $(0, \infty)$ *and strictly convex,* $\phi(1) = \phi'(1) = 0$, $\phi(0) < +\infty$, $\lim_{x \downarrow 0} \phi'(x) < 0$, $\lim_{x \to \infty} \phi(x)/x = +\infty$, $\lim_{x \to \infty} \phi'(x) > 0$, *and* $\lim_{x \to \infty} x\phi'(x)/\phi(x) < +\infty$.

For any $\phi \in \Phi_0$, let $\phi^\star(x) = \sup_{t \geq 0 : \phi(t) < +\infty}(tx - \phi(t))$ denote its convex conjugate and let $\psi(x) = \phi^\star(x) - x$. Define

$$\mathcal{E} = \{f : \mathcal{U} \to \mathbb{R} \text{ for which } \mathbb{E}^{F_*}[\psi(c|f(U)|)] < \infty \text{ for all } c > 0\}\,.$$

The class $\mathcal{E}$ is an Orlicz class (see Appendix D). For common choices of ϕ, we have

$$\begin{aligned}
&\mathcal{E} = \{f : \mathcal{U} \to \mathbb{R} : \mathbb{E}^{F_*}[e^{c|f(U)|}] < \infty \text{ for all } c > 0\} && \text{for KL divergence,} \\
&\mathcal{E} = \{f : \mathcal{U} \to \mathbb{R} : \mathbb{E}^{F_*}[f(U)^2] < \infty\} && \text{for } \chi^2 \text{ divergence, and} \\
&\mathcal{E} = \{f : \mathcal{U} \to \mathbb{R} : \mathbb{E}^{F_*}[|f(U)|^q] < \infty\} && \text{for } L^p \text{ divergence } (p^{-1} + q^{-1} = 1).
\end{aligned}$$

Let $g = (g_1, g_2, g_3, g_4)$ denote the vector formed by stacking each of the moment functions from (1a)–(1d). Our key regularity condition is the following assumption:

Assumption Φ (i) $\phi \in \Phi_0$.

(ii) $k(\,\cdot\,,\theta,\gamma)$ and each entry of $g(\,\cdot\,,\theta,\gamma)$ belong to $\mathcal{E}$ for each $\theta \in \Theta$ and $\gamma \in \Gamma$.

We briefly comment on Assumption Φ before proceeding. The ϕ functions inducing KL, χ^2, and L^p divergence all belong to Φ_0. For KL divergence, the class $\mathcal{E}$ contains of bounded functions (e.g. indicator functions) and functions that are additively separable in U provided F_* has tails that decay super-exponentially (e.g. Gaussian but not Gumbel). It therefore excludes several moment functions in the above examples under conventional parametric assumptions. Using χ^2 or L^p neighborhoods imposes much weaker requirements on the moment functions, requiring only finite second or qth moments, respectively.

The following result justifies formulating the criterion functions as finite-dimensional convex programs. Let $d = \sum_{i=1}^4 d_i$ where d_i is the dimension of g_i, let $\Lambda := \mathbb{R}_+^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}_+^{d_3} \times \mathbb{R}^{d_4}$, and let λ_{12} denote the first $d_1 + d_2$ elements of λ.

Proposition 2.1 *Let Assumption Φ hold. If there exists a distribution $F \in \mathcal{N}_\delta$ such that the moment conditions (1) hold at (θ, γ, P), then*

$$\underline{K}_\delta(\theta;\gamma,P) = \sup_{\eta>0,\zeta\in\mathbb{R},\lambda\in\Lambda} -\eta\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,\theta,\gamma)+\zeta+\lambda' g(U,\theta,\gamma)}{-\eta}\right)\right] - \eta\delta - \zeta - \lambda_{12}' P\,, \tag{13}$$

$$\overline{K}_\delta(\theta;\gamma,P) = \inf_{\eta>0,\zeta\in\mathbb{R},\lambda\in\Lambda} \eta\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,\theta,\gamma)-\zeta-\lambda' g(U,\theta,\gamma)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda_{12}' P\,. \tag{14}$$

Moreover, if the value of problem (13) is $+\infty$ (equivalently, if the value of problem (14) is $-\infty$), then there is no distribution in $\mathcal{N}_\delta$ under which (1) holds at (θ, γ, P).

The problems on the right-hand side of (13) and (14) are convex optimizations with respect to (η, ζ, λ). The parameter η is the Lagrange multiplier for the constraint $D_\phi(F\|F_*) \le \delta$. Similarly, λ collects the Lagrange multipliers for the moment in/equalities (1a)–(1d). These are unconstrained if they correspond to equality restrictions and non-negative if they correspond to inequality restrictions. Finally, ζ is the Lagrange multiplier for the constraint $\int \mathrm{d}F = 1$, thereby ensuring that the optimization is over probability measures.

It is worth highlighting some special cases of the dual formulation. For KL neighborhoods, $\phi^\star(x) = e^x - 1$ and the multiplier ζ may be solved out explicitly, leading to

$$\underline{K}_\delta(\theta;\gamma,P) = \sup_{\eta>0,\lambda\in\Lambda} -\eta \log \mathbb{E}^{F_*}\left[e^{-(k(U,\theta,\gamma)+\lambda' g(U,\theta,\gamma))/\eta}\right] - \eta\delta - \lambda_{12}' P\,,$$

$$\overline{K}_\delta(\theta;\gamma,P) = \inf_{\eta>0,\lambda\in\Lambda} \eta \log \mathbb{E}^{F_*}\left[e^{(k(U,\theta,\gamma)-\lambda' g(U,\theta,\gamma))/\eta}\right] + \eta\delta + \lambda_{12}' P\,.$$

Another special case is when k does not depend on u, i.e. $k(u,\theta,\gamma) = k(\theta,\gamma)$. In this case, the right-hand sides of (13) and (14) reduce to

$$\underline{K}_\delta(\theta;\gamma,P) = \left[\begin{array}{l} k(\theta,\gamma) \\ +\infty \end{array}\right. , \quad \overline{K}_\delta(\theta;\gamma,P) = \left[\begin{array}{ll} k(\theta,\gamma) & \text{if } \Delta^\star(\theta;\gamma,P) \le \delta, \\ -\infty & \text{if } \Delta^\star(\theta;\gamma,P) > \delta, \end{array}\right. \tag{15}$$

where

$$\Delta^\star(\theta;\gamma,P) = \sup_{\zeta\in\mathbb{R},\lambda\in\Lambda} -\mathbb{E}^{F_*}\left[\phi^\star(-\zeta-\lambda' g(U,\theta,\gamma))\right] - \zeta - \lambda_{12}' P\,. \tag{16}$$

The program $\Delta^\star$ is the dual formulation of the minimum-divergence problem

$$\Delta(\theta;\gamma,P) = \inf_F D_\phi(F\|F_*) \quad \text{subject to (1) holding at } (\theta,\gamma,P)\,. \tag{17}$$

The dual formulation is valid under Assumption Φ (see Appendix F.3 for a formal justification). For KL divergence, ζ may be solved out in closed form to obtain

$$\Delta^\star(\theta;\gamma,P) = \sup_{\lambda\in\Lambda} -\log \mathbb{E}^{F_*}\left[e^{-\lambda' g(U,\theta,\gamma)}\right] - \lambda_{12}' P\,.$$

An important feature of our approach is that the optimization problems (13), (14), and (16) are convex and their dimension does not increase with the neighborhood size δ. This feature is not shared by other seemingly natural approaches, such as using expanding parametric families (e.g. mixtures or other sieves) to flexibly model F. Importantly, our procedure recovers sharp bounds on the nonparametrically identified set of counterfactuals as δ becomes large.

2.4 Nonparametrically Identified Sets of Counterfactuals

We define the nonparametrically identified set of counterfactuals as

$$\mathcal{K} = \left\{\mathbb{E}^F[k(U,\theta,\gamma_0)] : (1) \text{ holds at } (\theta,\gamma_0,P_0),\ \theta\in\Theta,\ F\in\mathcal{F}_\theta\right\},$$

where

$$\mathcal{F}_\theta = \{F\in\mathcal{F} : \mathbb{E}^F[g(U,\theta,\gamma_0)] \text{ is finite and } F \ll \mu\}\,.$$

The set $\mathcal{F}_\theta$ consists of all probability measures on $\mathcal{U}$ that are absolutely continuous with respect to some σ-finite dominating measure μ and for which the moments in (1) are finite at θ. We impose existence of a density with respect to μ as it is often a structural assumption used, e.g., to avoid ties in CCPs or to establish existence of equilibria. Importantly, the class $\mathcal{F}_\theta$ can contain distributions not in $\mathcal{N}_\delta$ for any finite δ. It is therefore reasonable to ask: in confining ourselves to neighborhoods of the form $\mathcal{N}_\delta$, do we throw away other distributions that can yield smaller or larger values of the counterfactual? As we shall see, the answer is "no" provided F_* satisfies a type of full-support condition.

We require two other regularity conditions to establish the result. First, we say that k is "μ-essentially bounded" if $|k(\cdot,\theta,\gamma_0)|$ has finite μ-essential supremum[9] for each $\theta \in \Theta$. This condition is trivially satisfied when $k(u,\theta,\gamma_0)$ is a bounded function of u for each θ, which is true of the k functions representing counterfactual CCPs and change in average welfare in Examples 2.2 and 2.3. Note, however, that we do not require any of $g_1,\ldots,g_4$ to be bounded. Therefore, one may reparametrize models with unbounded k (e.g. Example 2.1) by setting $\tilde{\theta} = (\theta,\kappa)$, appending $k(U,\theta,\gamma_0) - \kappa$ as an element of g_4, and setting $k(U,\tilde{\theta},\gamma_0) = \kappa$.

The second is a Slater-type constraint qualification condition. Let 0_{d_i} denote a $d_i \times 1$ vector of zeros, $\mathcal{C} = \mathbb{R}^{d_1}_+ \times \{0_{d_2}\} \times \mathbb{R}^{d_3}_+ \times \{0_{d_4}\}$, $\mathcal{G}(\theta,\gamma) = \{\mathbb{E}^F[g(U,\theta,\gamma)] : D_\phi(F\|F_*) < \infty\}$, $\vec{P} = (P, 0_{d_3+d_4})$, and let $\mathrm{ri}(A)$ denote the relative interior of $A \subseteq \mathbb{R}^n$.

Definition 2.2 Condition S *holds at* (θ,γ,P) *if* $\vec{P} \in \mathrm{ri}(\mathcal{G}(\theta,\gamma) + \mathcal{C})$.

In models without inequality restrictions, Condition S requires that $\vec{P} \in \mathrm{ri}(\mathcal{G}(\theta,\gamma))$. For models with inequality conditions only, we require that there exists a distribution such that all inequalities are strict. Using "relative interior" instead of "interior" allows for moment functions that may be linearly dependent at certain values of θ.

Let $\Theta_I = \{\theta \in \Theta : (1) \text{ holds for some } F \in \mathcal{F}_\theta\}$ denote the nonparametrically identified set of structural parameters.

Theorem 2.1 *Let Assumption Φ hold, let Condition S hold at (θ,γ_0,P_0) for all $\theta \in \Theta_I$, let μ and F_* be mutually absolutely continuous, and let k be μ-essentially bounded. Then:*

$$\lim_{\delta\to\infty} \underline{\kappa}_\delta = \inf \mathcal{K}\,, \qquad\qquad \lim_{\delta\to\infty} \overline{\kappa}_\delta = \sup \mathcal{K}\,.$$

If μ is Lebesgue measure—which it often is in applications—then Theorem 2.1 implies that choosing F_* with strictly positive density over $\mathcal{U}$ ensures that $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ will recover the limits of $\mathcal{K}$ as δ gets large. Aside from sensitivity analyses, our procedure can therefore be used to compute bounds on $\mathcal{K}$ by setting the neighborhood size δ to be large but finite.

Remark 2.4 In Appendix B we characterize $\inf \mathcal{K}$ and $\sup \mathcal{K}$ using convex duality. Unlike Proposition 2.1, the dual representations are non-smooth, non-convex min-max and max-min problems which are computationally challenging, especially when u is multivariate.[10] Adding a ϕ-divergence constraint regularizes the dual formulation, resulting in the criterion functions $\underline{K}_\delta$ and $\overline{K}_\delta$ defined by smooth, convex optimization problems. In particular, the dimension of U does not play a role in the computational complexity of our procedure (though it may when expectations are computed numerically, as discussed in Section 3.1).

[9]The μ-essential supremum of $f : \mathcal{U} \to \mathbb{R}$ is $\mu\text{-ess}\sup f = \inf\{c : \mu(\{u : f(u) > c\} = 0)\}$.

[10]Similar representations are obtained in Ekeland et al. (2010) and Li (2018).

2.5 Estimation

The preceding discussion provides a roadmap for estimating the smallest and largest counterfactuals $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ given consistent first-stage estimators $(\hat{P}, \hat{\gamma})$ of (P_0, γ_0). Analogously to their population counterparts, estimators $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are obtained by optimizing sample criterion functions with respect to θ:

$$\hat{\underline{\kappa}}_\delta = \inf_{\theta \in \Theta} \hat{\underline{K}}_\delta(\theta)\,, \qquad \hat{\overline{\kappa}}_\delta = \sup_{\theta \in \Theta} \hat{\overline{K}}_\delta(\theta)\,,$$

where the sample criterion functions are

$$\hat{\underline{K}}_\delta(\theta) = \begin{bmatrix} \underline{K}_\delta(\theta; \hat{\gamma}, \hat{P}) \\ +\infty \end{bmatrix}, \quad \hat{\overline{K}}_\delta(\theta) = \begin{bmatrix} \overline{K}_\delta(\theta; \hat{\gamma}, \hat{P}) & \text{if } \Delta^\star(\theta; \hat{\gamma}, \hat{P}) < \delta, \\ -\infty & \text{if } \Delta^\star(\theta; \hat{\gamma}, \hat{P}) \geq \delta, \end{bmatrix}$$

with $\overline{K}_\delta(\theta; \hat{\gamma}, \hat{P})$ and $\overline{K}_\delta(\theta; \hat{\gamma}, \hat{P})$ denoting the values of the convex optimization problems from Proposition 2.1 evaluated at $(\hat{\gamma}, \hat{P})$, and $\Delta^\star(\theta; \hat{\gamma}, \hat{P})$ denoting the value of the optimization problem (16) evaluated at $(\hat{\gamma}, \hat{P})$. If $k(u, \theta, \gamma) = k(\theta, \gamma)$, the criterion functions simplify:

$$\hat{\underline{K}}_\delta(\theta) = \begin{bmatrix} k(\theta, \hat{\gamma}) \\ +\infty \end{bmatrix}, \quad \hat{\overline{K}}_\delta(\theta) = \begin{bmatrix} k(\theta, \hat{\gamma}) & \text{if } \Delta^\star(\theta; \hat{\gamma}, \hat{P}) < \delta, \\ -\infty & \text{if } \Delta^\star(\theta; \hat{\gamma}, \hat{P}) \geq \delta. \end{bmatrix}$$

In Section 6 we establish consistency of the estimators $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$, derive their joint asymptotic distribution, and present two procedures for inference on $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$.

3 Practical Considerations

3.1 Computation

There are three aspects to computation: (i) computing the expectations with respect to F_* in the objective functions, (ii) solving the inner optimization problems over Lagrange multipliers, and (iii) solving the outer optimization problems over θ.

The expectations in the objective functions (13), (14), and (16) are available in closed-form for certain models, ϕ-divergences, and distributions F_*,[11] in which case the dimension of U does not play a role in the computational complexity of our procedure. Outside of these special cases, the expectations in the objective functions will need to be computed numerically. Here the dimension of U will play a role in terms of determining how many quadrature points or Monte Carlo draws are needed to control any numerical approximation error. In practice, we used randomized quasi-Monte Carlo methods, specifically scrambled Halton sequences using

[11] An earlier draft worked through the closed-form solution for a discrete game of complete information with Gaussian payoff shocks and KL neighborhoods (see `https://arxiv.org/abs/1904.00989v2`).

the algorithm of Owen (2017), which can yield improvements over conventional Monte Carlo methods especially when U is of moderate or high dimension (see, e.g., Lemieux (2009)).

The inner optimization with respect to Lagrange multipliers can be solved rapidly: it is convex and gradients and Hessians are available in closed-form. The envelope theorem can be used to derive gradients for the outer optimization when k and the moment functions are continuously differentiable in θ. Our procedures were all implemented in Julia with the inner and outer optimizations solved using KNITRO. As with parameter estimation in nonlinear structural models, the outer optimization with respect to θ is typically non-convex. In applications, we iteratively applied a multi-start procedure in the outer optimization in an attempt to avoid local optima. Computation times are reported in the applications below.

3.2 MPEC Implementation

In this subsection, we propose a MPEC version of our procedure to simplify computation in models with endogenous parameters defined by equilibrium conditions (e.g. value functions defined by Bellman equations). In DDC models as described Example 2.3, this approach reduces the dimension of the inner optimization by twice the cardinality of the statespace.

Suppose we may partition $\theta = (\theta_s, \theta_e)$ and $g_4 = (g_{4s}, g_{4e})$ where θ_s are "deep" structural parameters and θ_e are "endogenous" parameters that are defined implicitly by the subset of moment conditions corresponding to g_{4e}. That is, for any (θ_s, γ, F), the parameter $\theta_e = \theta_e(\theta_s, \gamma, F)$ solves

$$\mathbb{E}^F[g_{4e}(U, (\theta_s, \theta_e), \gamma)] = 0\,.$$

For instance, in Example 2.3 we have $\theta_s = (\theta_\pi, \beta)$, $\theta_e = (v, \tilde{v})$, and g_{4e} is the vector of moment functions in display (6). While our procedure can be implemented as described in Section 2, this does not take advantage of the fact that the subvector θ_e is defined implicitly.

To leverage this additional structure, consider the subset of moments excluding g_{4e}:

$$\begin{aligned} &\mathbb{E}^F[g_1(U,\theta,\gamma_0)] \le P_{10}, \quad \mathbb{E}^F[g_2(U,\theta,\gamma_0)] = P_{20}, \\ &\mathbb{E}^F[g_3(U,\theta,\gamma_0)] \le 0, \qquad \mathbb{E}^F[g_{4s}(U,\theta,\gamma_0)] = 0, \end{aligned} \tag{18}$$

and define the criterion functions using only these moments:

$$\underline{K}^s_\delta(\theta;\gamma,P) = \inf_{F\in\mathcal{N}_\delta} \mathbb{E}^F[k(U,\theta,\gamma)] \quad \text{subject to (18) holding at } (\theta,\gamma,P)\,, \tag{19}$$

$$\overline{K}^s_\delta(\theta;\gamma,P) = \sup_{F\in\mathcal{N}_\delta} \mathbb{E}^F[k(U,\theta,\gamma)] \quad \text{subject to (18) holding at } (\theta,\gamma,P)\,. \tag{20}$$

Under the conditions of Proposition 2.1, these criterion functions admit a dual formulation:

$$\underline{K}^s_\delta(\theta;\gamma,P) = \sup_{\eta>0,\zeta\in\mathbb{R},\lambda\in\Lambda_s} -\eta\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,\theta,\gamma)+\zeta+\lambda' g_s(U,\theta,\gamma)}{-\eta}\right)\right] - \eta\delta - \zeta - \lambda_{12}' P\,, \tag{21}$$

$$\overline{K}^s_\delta(\theta;\gamma,P) = \inf_{\eta>0,\zeta\in\mathbb{R},\lambda\in\Lambda_s} \eta\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,\theta,\gamma)-\zeta-\lambda' g_s(U,\theta,\gamma)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda_{12}' P\,, \tag{22}$$

with $g_s = (g_1, g_2, g_3, g_{4s})$ and $\Lambda_s = \mathbb{R}^{d_1}_+ \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_3}_+ \times \mathbb{R}^{d_{4s}}$ with $d_{4s} = \dim(g_{4s})$. These criterion functions simplify analogously to (15) when k does not depend on u, with the minimum divergence problem $\Delta^\star$ now involving only the g_s moments. The remaining moment conditions corresponding to g_{4e} are appended as constraints in the outer optimization. The constraints are evaluated under the distributions that solve problems (19) and (20), which we denote $\underline{F}_{\delta,\theta}$ and $\overline{F}_{\delta,\theta}$.[12] We first justify this approach and then show how to construct $\underline{F}_{\delta,\theta}$ and $\overline{F}_{\delta,\theta}$.

Proposition 3.1 *Let Assumption Φ hold. Then:*

$$\inf_{\theta\in\Theta} \underline{K}_\delta(\theta;\gamma,P) \equiv \inf_{\theta\in\Theta} \underline{K}^s_\delta(\theta;\gamma,P) \ \ \textit{subject to} \ \ \mathbb{E}^{\underline{F}_{\delta,\theta}}[g_{4e}(U,\theta,\gamma)] = 0\,,$$
$$\sup_{\theta\in\Theta} \overline{K}_\delta(\theta;\gamma,P) \equiv \sup_{\theta\in\Theta} \overline{K}^s_\delta(\theta;\gamma,P) \ \ \textit{subject to} \ \ \mathbb{E}^{\overline{F}_{\delta,\theta}}[g_{4e}(U,\theta,\gamma)] = 0\,.$$

Proposition 3.1 justifies our MPEC implementation, showing that it yields the same minimizing and maximizing values of the counterfactual as the "full" implementation described in Section 2. The expectations in the constraints may be expressed as

$$\mathbb{E}^{\underline{F}_{\delta,\theta}}[g_{4e}(U,\theta,\gamma)] = \mathbb{E}^{F_*}[\underline{m}_{\delta,\theta}(U) g_{4e}(U,\theta,\gamma)]\,,$$
$$\mathbb{E}^{\overline{F}_{\delta,\theta}}[g_{4e}(U,\theta,\gamma)] = \mathbb{E}^{F_*}[\overline{m}_{\delta,\theta}(U) g_{4e}(U,\theta,\gamma)]\,,$$

where $\underline{m}_{\delta,\theta}$ and $\overline{m}_{\delta,\theta}$ denote the Radon–Nikodym derivatives of $\underline{F}_{\delta,\theta}$ and $\overline{F}_{\delta,\theta}$ with respect to F_*. When k depends on u, we construct $\underline{m}_{\delta,\theta}$ and $\overline{m}_{\delta,\theta}$ from solutions to (21) and (22), say $(\underline{\eta},\underline{\zeta},\underline{\lambda})$ and $(\overline{\eta},\overline{\zeta},\overline{\lambda})$ (solutions exist under the regularity conditions below). With $\underline{\eta} > 0$, define

$$\underline{m}_{\delta,\theta}(u) = \dot{\phi}^\star\left(\frac{k(u,\theta,\gamma) + \underline{\zeta} + \underline{\lambda}' g_s(u,\theta,\gamma)}{-\underline{\eta}}\right)\,, \tag{23}$$

where $\dot{\phi}^\star(x) = \frac{d\phi^\star(x)}{dx}$. The function $\overline{m}_{\delta,\theta}(u)$ is constructed similarly, replacing $(\underline{\eta},\underline{\zeta},\underline{\lambda})$ in (23) by $(-\overline{\eta},-\overline{\zeta},-\overline{\lambda})$. For KL divergence the multiplier $\underline{\zeta}$ can be solved out explicitly, leading to

$$\underline{m}_{\delta,\theta}(u) = \frac{e^{(k(u,\theta,\gamma)+\underline{\lambda}' g_s(u,\theta,\gamma))/-\underline{\eta}}}{\mathbb{E}^{F_*}\left[e^{(k(u,\theta,\gamma)+\underline{\lambda}' g_s(u,\theta,\gamma))/-\underline{\eta}}\right]}\,.$$

[12] The distributions $\underline{F}_{\delta,\theta}$ and $\overline{F}_{\delta,\theta}$ also depend on γ and P, but we suppress this to simplify notation.

When $\underline{\eta} > 0$ the distribution solving (19) is unique and is defined by $\underline{m}_{\delta,\theta}(u)$.[13] We exclude the case $\underline{\eta} = 0$ as it is rare in practice.[14]

When k does not depend on u, $\underline{m}_{\delta,\theta}$ and $\overline{m}_{\delta,\theta}$ are constructed from solutions to the program $\Delta^\star(\theta;\gamma,P)$ defined in (16) using only the reduced set of moments g_s. Under the regularity conditions below, this program has a solution, say $(\underline{\zeta},\underline{\lambda})$, and we define

$$\underline{m}_{\delta,\theta}(u) = \overline{m}_{\delta,\theta}(u) = \dot{\phi}^\star\left(-\underline{\zeta} - \underline{\lambda}' g_s(u,\theta,\gamma)\right) . \tag{24}$$

For KL divergence the multiplier $\underline{\zeta}$ may be solved out explicitly, leading to

$$\underline{m}_{\delta,\theta}(u) = \overline{m}_{\delta,\theta}(u) = \frac{e^{-\underline{\lambda}' g_s(u,\theta,\gamma)}}{\mathbb{E}^{F_*}\left[e^{-\underline{\lambda}' g_s(u,\theta,\gamma)}\right]} .$$

Proposition 3.2 *Let Assumption Φ hold, let Condition S hold at (θ,γ,P), and let there exist a distribution F with $D(F\|F_*) < \delta$ under which conditions (18) hold at (θ,γ,P). Then: the distributions $\underline{F}_{\delta,\theta}$ and $\overline{F}_{\delta,\theta}$ induced by $\underline{m}_{\delta,\theta}$ and $\overline{m}_{\delta,\theta}$ solve (19) and (20).*

Example. In Table 2 we report computation times for the inner optimization problems for the DDC model and counterfactual in Section 5.4 of Norets and Tang (2014), which is based on Rust (1987), in which $|\mathcal{S}| = 90$ and $U = (U_0, U_1)$. For brevity, we report times for the inner problems for $\overline{K}_\delta$ and $\overline{K}_\delta^s$ for maximizing the counterfactual CCP in the highest mileage state only. We also report times for solving the inner problem for the minimum divergence problem $\Delta^\star$ and its MPEC analogue using g_s only. The MPEC implementation has 92 moments in g_s (90 moments for the CCPs plus two mean-zero normalizations for U_0 and U_1) for the inner optimization. The inner optimization in the full implementation involves an additional 180 moments representing the Bellman equations pre- and post-intervention. In this design, the inner optimization problems are solved at least 20 times faster for the MPEC implementation.[15]

3.3 Overidentification

Our theoretical results are developed assuming the researcher's model is correctly specified: there exists $\theta \in \Theta$ for which the population moment conditions (1) hold under F_*. However, in overidentified models (i.e., $d > d_\theta$) there might not exist $\theta \in \Theta$ for which the sample moment conditions (7) hold under F_*. We propose two options to deal with this situation.

[13] Uniqueness follows by strict convexity of ϕ because $D_\phi(F\|F_*) \le \delta$ must be binding at $\underline{F}_{\delta,\theta}$ if $\underline{\eta} > 0$.

[14] In this case, the distribution $\underline{F}_{\delta,\theta}$ will concentrate on the F_*-essential infimum of $k(\cdot,\theta,\gamma) + \underline{\lambda}' g_s(\cdot,\theta,\gamma)$. If k and/or elements of g_s are unbounded then this set will typically have zero F_* measure. In consequence, $\underline{\eta} = 0$ will not be a solution for any $\delta < +\infty$.

[15] The computations reported in Table 2 initialize the solver at $\eta = 1$, $\zeta = 0$, and $\lambda = 0$. When embedded in the outer optimization with respect to θ, the inner-optimization computation times are reduced significantly by initializing at the (η,ζ,λ) solving the inner problem at the previous value of θ.

Implementation	$\overline{K}_{0.01}$	$\overline{K}_{0.10}$	$\overline{K}_{1.00}$	$\Delta^\star$
MPEC (92 moments)	0.185	0.209	0.244	0.102
Full (272 moments)	4.444	12.028	33.250	3.585

Table 2: Computation times (seconds) for solving the inner optimization problems using the design in Section 5.4 of Norets and Tang (2014) at the true θ values. Expectations are computed numerically using $50,000$ scrambled Halton draws. Neighborhoods constrained by a hybrid of KL and χ^2 divergence (see Section 5). All computations are performed in Julia version 1.5.3 and KNITRO 12.2.0 on a 2.7GHz MacBook Pro with 16GB memory.

First, one may compute the smallest value of δ for which there exists $F \in \mathcal{N}_\delta$ consistent with the sample moment conditions (7) by solving

$$\hat{\delta} = \inf_{\theta \in \Theta} \Delta^\star(\theta; \hat{\gamma}, \hat{P}).$$

The bounds $[\hat{\underline{\kappa}}_\delta, \hat{\overline{\kappa}}_\delta]$ will be nonempty for $\delta > \hat{\delta}$. Under the conditions of Theorem 6.1, the value $\hat{\delta}$ will converge in probability to zero and for each $\delta > 0$ the bounds $[\hat{\underline{\kappa}}_\delta, \hat{\overline{\kappa}}_\delta]$ will be nonempty with probability approaching one.

For certain correctly specified but overidentified models it may be the case that $\hat{\delta} = +\infty$. This can occur when estimates $\hat{P}$ are inconsistent with certain restrictions of the model. For instance, it is natural to estimate CCPs nonparametrically using the empirical frequencies of choices in each state. If some choices aren't observed in certain states then their corresponding estimated CCPs will be zero even though the model-implied CCPs may be strictly positive.

In models with equality restrictions only for P (i.e., $P_0 \equiv P_{20}$), this issue can be avoided by using $\hat{P} = \mathbb{E}^{F_*}[g_2(U, \tilde{\theta}, \hat{\gamma})]$ where $\tilde{\theta}$ is an estimate of θ based on (7). This approach ensures $\hat{P}$ is compatible with the model and therefore that the bounds $[\hat{\underline{\kappa}}_\delta, \hat{\overline{\kappa}}_\delta]$ are nonempty for each δ. Under mild regularity conditions the estimator $\hat{P}$ will be consistent and asymptotically normal, so the consistency and inference results developed in Section 6 will continue to apply.

4 Interpreting the Neighborhood Size

As with any sensitivity analysis, interpreting the neighborhood size δ is important. In this section we first discuss some properties of ϕ-divergences and their implications for interpreting δ. We then present some specific methods for interpretation.

4.1 Invariance

A defining property of ϕ-divergences that they are invariant to invertible transformations. That is, if G and G_* denote the distribution of $T(U)$ when $U \sim F$ and $U \sim F_*$, respectively, for some invertible transformation T, then $D_\phi(F\|F_*) = D_\phi(G\|G_*)$.[16] An important consequence is that the neighborhood size has the same interpretation under a change in units for U. For instance, if one researcher writes a model using dollars as units with $U \sim F_*$ and another uses thousands of dollars as units with $U \sim G_*$ where $G_*(u) = F_*(10^{-3}u)$ then F is in $\mathcal{N}_\delta$ if and only if its rescaled counterpart G is in the corresponding δ-neighborhoods of G_*. A second consequence is that neighborhood size is invariant under location and scale transformations of F_* (e.g. $N(\mu, \Sigma)$ versus $N(0, I)$).

4.2 Relating Different ϕ-divergences

It is well known that ϕ-divergences are equivalent over vanishingly small neighborhoods.[17] While the bounds $[\underline{\kappa}_\delta, \overline{\kappa}_\delta]$ may depend on the choice of ϕ over non-vanishing neighborhoods, it is possible to formally relate the bounds induced by different ϕ functions.

Consider two functions, say ϕ_1 and ϕ_2. We write $\mathcal{N}_{\phi_1,\delta}$ and $\mathcal{N}_{\phi_2,\delta}$ for δ-neighborhoods from (10) induced by ϕ_1 and ϕ_2, respectively. Without loss of generality, order ϕ_1 and ϕ_2 so that $\lim_{x\to+\infty} \frac{\phi_1(x)}{\phi_2(x)} < \infty$. Under this ordering, the divergence induced by ϕ_2 is stronger than that induced by ϕ_1: $D_{\phi_2}(F\|F_*) < \infty$ implies $D_{\phi_1}(F\|F_*) < \infty$ (see Appendix D). Define

$$\bar{a} := \sup_{x>0, x\neq 1} \frac{\phi_1(x)}{\phi_2(x)}\,.$$

The quantity $\bar{a}$ is a measure of relative neighborhood size in the sense that $\mathcal{N}_{\phi_2,\delta} \subseteq \mathcal{N}_{\phi_1,\bar{a}\delta}$ holds for each $\delta > 0$ (see the proof of Proposition 4.1). Thus, δ-neighborhoods under ϕ_1 are $\bar{a}$ times as large as δ neighborhoods under ϕ_2. For instance, KL δ-neighborhoods are twice as large as χ^2 δ-neighborhoods. Finally, write $\underline{\kappa}_{\phi_1,\delta}$ and $\underline{\kappa}_{\phi_2,\delta}$ for the smallest counterfactual from (8) over $\mathcal{N}_{\phi_1,\delta}$ and $\mathcal{N}_{\phi_2,\delta}$. The values $\overline{\kappa}_{\phi_1,\delta}$ and $\overline{\kappa}_{\phi_2,\delta}$ are defined analogously.

Proposition 4.1 *Let $\phi_1, \phi_2 \in \Phi_0$ and let Assumption Φ(ii) hold for the space $\mathcal{E}$ corresponding to ϕ_1. Then: the inclusion $[\underline{\kappa}_{\phi_2,\delta}, \underline{\kappa}_{\phi_2,\delta}] \subseteq [\underline{\kappa}_{\phi_1,\bar{a}\delta}, \underline{\kappa}_{\phi_1,\bar{a}\delta}]$ holds for each $\delta > 0$.*

It follows from Proposition 4.1 that bounds that are wide under ϕ_2 must necessarily be wide under ϕ_1. Similarly, narrow bounds under ϕ_1 must also be narrow under ϕ_2. Note also that the ordering does not depend on the counterfactual function k.

4.3 "Least Favorable" Distributions

A useful feature of our approach is that the "least favorable" distributions (LFDs) that attain the smallest or largest values of the counterfactual can be recovered. The researcher may plot

[16] See, e.g., Liese and Vajda (1987). A more direct statement is in Qiao and Minematsu (2010), who also show invariance is unique to ϕ-divergences.

[17] See, e.g., Theorem 4.1 of Csiszár and Shields (2004).

the LFDs and compute other quantities of interest (e.g. correlations or welfare measures) under the LFDs to help interpret δ.

As in Section 3.2, we use Radon–Nikodym derivatives to construct the LFD. If k depends on u then the Radon–Nikodym derivative of the distribution solving (11) at θ is analogous to (23), namely

$$\underline{m}_{\delta,\theta}(u) = \dot{\phi}^{\star}\left(\frac{k(u,\theta,\gamma)+\underline{\zeta}+\underline{\lambda}'g(u,\theta,\gamma)}{-\underline{\eta}}\right), \tag{25}$$

where $(\underline{\eta},\underline{\zeta},\underline{\lambda})$ is any solution to (13) with $\underline{\eta} > 0$ (as in Section 3.2, we exclude the case $\underline{\eta} = 0$ as it is rare in practice). The distribution solving the maximization problem (12) at θ is obtained similarly, replacing $(\underline{\eta},\underline{\zeta},\underline{\lambda})$ in (25) by $(-\overline{\eta},-\overline{\zeta},-\overline{\lambda})$ where $(\overline{\eta},\overline{\zeta},\overline{\lambda})$ solves (14). The distributions solving (11) and (12) are unique when $\underline{\eta} > 0$ and $\overline{\eta} > 0$, respectively (see Section 3.2). If k does not depend on u then we set

$$\underline{m}_{\delta,\theta}(u) = \overline{m}_{\delta,\theta}(u) = \dot{\phi}^{\star}\left(-\underline{\zeta}-\underline{\lambda}'g(u,\theta,\gamma)\right) \tag{26}$$

where $(\underline{\zeta},\underline{\lambda})$ solve the program (16). While there may exist multiple distributions solving (11) and (12) at θ in this case, the distribution induced by (26) has smallest ϕ-divergence relative to F_*. The justification of these LFD constructions follows similarly to Proposition 3.2.

4.4 Viewing Neighborhood Size through the Lens of the Model

Our second method for interpreting the neighborhood size is based on measuring variation in the moments at the solutions to the optimization problems (8) and (9) relative to their values under the researcher's specification F_*.

Under the regularity conditions in Section 6, the values $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are attained by minimizing and maximizing $\underline{K}_\delta(\theta;\gamma_0,P_0)$ and $\overline{K}_\delta(\theta;\gamma_0,P_0)$ with respect to θ, and the sets of minimizing and maximizing values of θ, denoted $\underline{\Theta}_\delta$ and $\overline{\Theta}_\delta$, are nonempty. While any $\theta \in \underline{\Theta}_\delta \cup \overline{\Theta}_\delta$ satisfies the moment conditions (1) under the corresponding LFD, it will typically *not* do so under F_*. Our second, model-based interpretation for the neighborhood size is to consider the maximum degree to which the moments at $\theta \in \underline{\Theta}_\delta \cup \overline{\Theta}_\delta$ violate (1) under F_*:

$$\begin{aligned} size(\delta) = \sup_{\theta\in\underline{\Theta}_\delta\cup\overline{\Theta}_\delta} \max\Big\{ & \Big\| \left(\mathbb{E}^{F_*}[g_1(U,\theta,\gamma_0)] - P_{10}\right)_+ \Big\|_\infty, \Big\|\mathbb{E}^{F_*}[g_1(U,\theta,\gamma_0)] - P_{20}\Big\|_\infty, \\ & \Big\| \left(\mathbb{E}^{F_*}[g_3(U,\theta,\gamma_0)]\right)_+ \Big\|_\infty, \Big\|\mathbb{E}^{F_*}[g_4(U,\theta,\gamma_0)]\Big\|_\infty \Big\}, \end{aligned}$$

where $(v)_+ = ((v_1 \vee 0),\ldots,(v_d \vee 0))$ for a vector $v \in \mathbb{R}^d$. This measure is informative about the extent to which the moment conditions (1) discipline the model's predictions about κ once F is allowed to vary over $\mathcal{N}_\delta$. Small values of $size(\delta)$ indicate that the moment conditions (1) exert little influence on the counterfactuals: the LFDs supporting $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ distort F_* in a way that moves the counterfactual but barely moves the moments. Conversely, large values of $size(\delta)$ indicate that distortions required to increase or decrease the counterfactual also have a material impact on the moments. This measure can be computed in practice by replacing (P_0, γ_0) by estimators $(\hat{P}, \hat{\gamma})$ and the sets $\underline{\Theta}_\delta$ and $\overline{\Theta}_\delta$ by either the minimizers and maximizers of the sample criterions or by the estimators of $\underline{\Theta}_\delta$ and $\overline{\Theta}_\delta$ described in Section 6.

5 Empirical Applications

5.1 Marital College Premium

Chiappori, Salanié, and Weiss (2017, CSW hereafter) study the evolution of marital returns to education in the US using a matching model with transferable utility (Choo and Siow, 2006). CSW find that the marital college premium—the difference in social surplus that college- and non-college-educated individuals derive from the marriage market—increased significantly across cohorts of women in the US in the mid to late 20th century, which they attribute to increasing returns to parental investment in children's human capital. As is conventional following Dagsvik (2000) and Choo and Siow (2006), CSW assume that the random components of individuals' marital preferences are i.i.d. Gumbel. While this assumption lends analytic tractability, the distribution of preference shocks is not nonparametrically identified in CSW's model. We therefore examine the sensitivity of their findings to this conventional assumption.

Our analysis reveals several insights about CSW's findings and, more generally, about welfare measures in matching models following Choo and Siow (2006). First, estimates of the marital college premium are highly sensitive to the i.i.d. Gumbel assumption and some of CSW's findings cannot be maintained under slight departures from this assumption. Interestingly, this sensitivity arises primarily at the estimation stage: premiums have narrow nonparametrically identified sets at any *fixed* value of parameters, yet relaxing the parametric assumption slightly allows significant variation in parameters and, consequently, wide bounds on premiums. In the framework of Choo and Siow (2006), preference parameters are just-identified under for any given distribution of shocks. Therefore, overidentifying restrictions on parameters and/or shape restrictions on the distribution are required to tighten the bounds. We explore the role of one such shape restriction, namely exchangeability, in tightening the bounds. Overall, we find weaker, but more robust evidence consistent with CSW's theme of increasing marital returns to higher education.

Model. The model follows Choo and Siow (2006). Agents are male or female and one of J types. Each agent chooses to be unmatched or to match with one partner of the opposite gender. A type-a man i receives utility ε_{ia0} if unmatched and $z_{ab} + \varepsilon_{iab}$ if he matches with a type-b female; a type-b female i' receives utility $e_{i'0b}$ if unmatched and $t_{ab} + e_{i'ab}$ if she matches with a type-a male. The parameters $(z_{ab}, t_{ab})_{a,b=1}^J$ represent the common deterministic component of marital preferences whereas the latent (to the econometrician) shocks $(\varepsilon_{ia0}, \ldots, \varepsilon_{iaJ})$ and $(e_{i'0b}, \ldots, e_{i'Jb})$ represent the idiosyncratic components. Shocks are i.i.d. across agents (conditional on own type and gender) with mean zero.

The stable matching problem can be reduced to standard individual-level discrete choice problems (see, e.g., Propositions 1 and 2 of CSW for a formal statement), so this model maps into Example 2.1. CSW estimate the model using data from the American Community Survey. They form 28 cohorts indexed by female birth year from 1941 (cohort 1) to 1968 (cohort 28) and treat each as an independent marriage market. We use a c superscript in what follows to denote cohort-c quantities. For each cohort c, CSW estimate the parameters $(t^c_{ab})^J_{a,b=1}$ and $(z^c_{ab})^J_{a,b=1}$ by matching estimated and model-implied match probabilities for females and males, respectively, assuming the $(e_{0b}, \ldots, e_{Jb})^J_{b=1}$ and $(\varepsilon_{a0}, \ldots, \varepsilon_{aJ})^J_{a=1}$ are i.i.d. Gumbel (F_* hereafter).

The type b to b' marital education premium for cohort-c females is

$$\kappa^c := \mathbb{E}^F\left[\max_{a=1,\ldots,J}\left(t^c_{ab'} + e_{ab'}\right) \vee e_{0b'}\right] - \mathbb{E}^F\left[\max_{a=1,\ldots,J}\left(t^c_{ab} + e_{ab}\right) \vee e_{0b}\right], \qquad (27)$$

which corresponds to the welfare measure ΔW in Example 2.1. CSW estimate the premium by evaluating (27) at their estimates of $(t^c_{ab}, t^c_{ab'})^J_{a=1}$ under F_*.

Implementation. We focus on CSW's estimates for marriages between white men and women for which there are $J = 5$ ordered types, namely "high-school dropouts" (type 1), "high-school graduates", "some college", "college graduate", and "college-plus" (type 5).

Our first implementation lets the distribution of $(e_{0b}, \ldots, e_{Jb})$ vary by cohort but not by own type b.[18] For each cohort c, we therefore let $(e_{ab})^J_{a=0} =_d U \sim F^c$ for all b. The parameters $(t^c_{ab})^J_{a=1}$ are just-identified from the match probabilities for cohort-c, type-b women, say $(\Pr^c(a|b))^J_{a=1}$, under any given distribution F^c (see, e.g., Galichon and Salanié (2020)). Similarly, $(z^c_{ab})^J_{a,b=1}$ is just-identified from the cohort-c match probabilities for men under any given distribution of the male preference shocks. We therefore only need to enforce the moment conditions that involve the parameters in (27), namely $\theta^c := (t^c_{ab}, t^c_{ab'})^J_{a=1}$, as the remaining parameters can be chosen to fit the other match probabilities under the resulting LFD. We therefore form g_2 as in Example 2.1 to explain the type b and b' match probabilities $P^c_{20} := (\Pr^c(a|b), \Pr^c(a|b'))^J_{a=1}$ and use CSW's estimated match probabilities as our estimate $\hat{P}^c_2$ of P^c_{20}. We also normalize the marginal distribution of each shock to have mean zero and the same variance as under F_*, forming $g_4(U, \theta) = (U_j, U_j^2 - \pi^2/6)^J_{j=0}$. The scale normalization ensures that the nonparametrically identified set for the premium is bounded at any fixed θ^c. As $J = 5$, there are 22 moments (for 10 match probabilities and 12 location/scale normalizations) and 10 parameters in this implementation.

[18] In view of the just-identification results below, we would obtain the same results if the distribution was homogeneous across cohorts. Allowing for heterogeneity in own-type would result in wider bounds.

Our second implementation imposes invariance of the distribution F^c under rotations and reflections of potential spouse types ("dihedral exchangeability"; see Appendix A.3). Under this shape restriction, match probabilities depend on the parameters $(t^c_{ab})^J_{ab=1}$ but not the labeling of potential spouse types (though they may depend to some extent on the ordering of types).[19] That is, F^c must explain $(\Pr^c(a|b))^J_{a=1}$ when $(e_{\varpi(0)b}, \ldots, e_{\varpi(J)b}) \sim F^c$ for each dihedral permutation ϖ of $\{0, \ldots, J\}$ (of which there are 12 when $J = 5$), and similarly for type b'. The moment conditions from our first implementation are therefore duplicated across all 12 permutations, so there are 264 moments and 10 parameters in this implementation.[20] By symmetry, however, it suffices to form g_2 and g_4 using the averages of the 22 moments across the 12 permutations rather than all 264 moments separately (see Remark A.1).

Assumption Φ(ii) fails for KL divergence for this model but holds for χ^2 and L^p ($p < \infty$) divergence. One possibly unappealing feature of χ^2 divergence, however, is that the LFDs can assign zero mass to certain regions. We therefore define neighborhoods using a hybrid of KL and χ^2 divergence induced by

$$\phi(x) = \begin{cases} x \log x - x + 1 & \text{if } x \le e, \\ \frac{1}{2e}(x-e)^2 + (x-e) + 1 & \text{if } x > e. \end{cases}$$

This hybrid divergence, like KL, ensures the LFDs are everywhere positive and only requires finite second moments for Assumption Φ(ii) to hold, which is indeed the case. As a robustness check, we repeated our analysis with neighborhoods constrained by χ^2 and L^4 divergences. Overall, our findings are not sensitive to the choice of ϕ (see Appendix C.1 for a discussion).

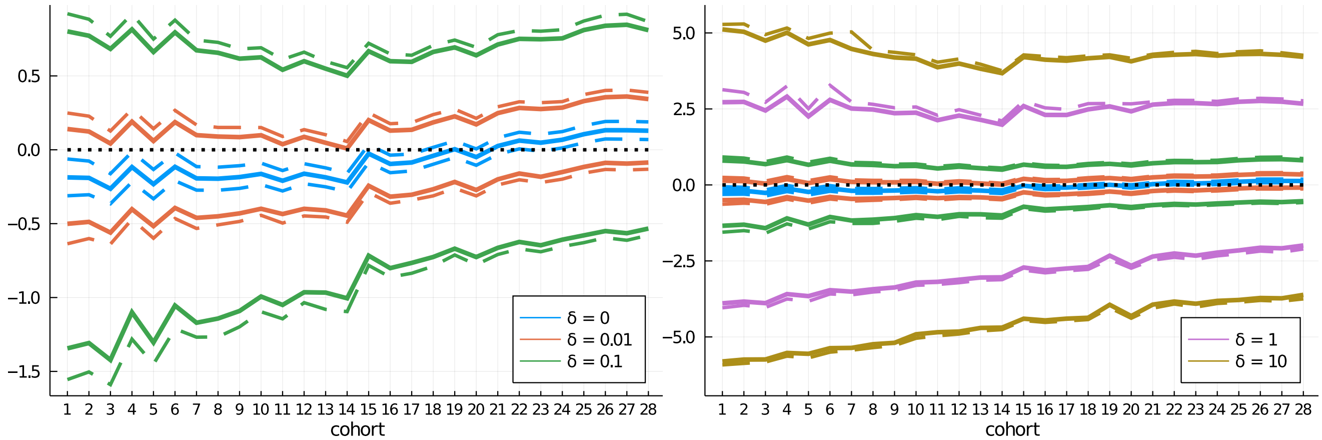
All computations are performed as described in Section 3.1. Our first implementation uses $50,000$ scrambled Halton draws to compute the expectations. Our second uses $15,000$ draws which are concatenated over the 12 permutations (see Remark A.2), for a total of $180,000$ draws. Computation times are reported in Appendix C.1.

Findings. We focus two findings from CSW, namely (i) that the "some college" to "college graduate" premium for white women went from being significantly negative in early cohorts to significantly positive for late cohorts, and (ii) that the "college graduate" to "college-plus" premium for white women, while negative, increased significantly across cohorts.

CSW's estimates of the "some college" to "college graduate" (SC to CG) premium under F_* are plotted across cohorts in blue in Figure 1a with (pointwise) 95% confidence intervals (cf. Panel C of Figure 21 in CSW). We relax F_* and plot estimates and confidence sets for $\underline{\kappa}^c_\delta$ and $\overline{\kappa}^c_\delta$ for our first implementation in Figure 1, beginning at $\delta = 0.01$ and increasing δ by factors of

[19] Allowing dependence on ordering seems desirable in this setting where types are naturally ordered. This dependence can be shut down by imposing (full) exchangeability on F^c. Doing so was computationally infeasible with $J = 5$ as it required computing the moments under all $(J + 1)! = 720$ permutations across a further 720 permutations of many draws (see Remark A.2), though it is feasible with smaller J.

[20] If F^c is dihedrally exchangeable then these $2(J + 1) \times 2J$ restrictions are just-identifying for $(t^c_{ab}, t^c_{ab'})^J_{a=1}$.



(a) Small δ-neighborhoods.

(b) Large δ-neighborhoods.

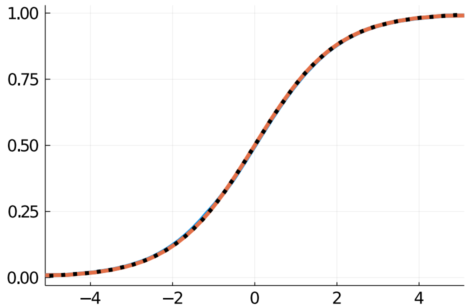
Figure 1: Sensitivity analysis for the SC to CG premium κ^c for white women for our first implementation. CSW's estimates under F_* are shown in blue ($\delta = 0$). Estimates of $\underline{\kappa}_\delta^c$ and $\overline{\kappa}_\delta^c$ are shown as solid lines; dashed lines are (pointwise) 95% confidence sets for $\underline{\kappa}_\delta^c$ and $\overline{\kappa}_\delta^c$.

10.[21] With $\delta = 0.01$, the estimates $\hat{\underline{\kappa}}_\delta^c$ and $\hat{\overline{\kappa}}_\delta^c$ lie uniformly below and above zero across cohorts (see Figure 1a). Therefore, it is impossible to say how the sign of the premium has changed over time when the i.i.d. Gumbel assumption is relaxed even slightly without further restrictions on F^c or θ^c. Figure 1b shows that the bounds become uninformatively wide for large δ. There are some interesting asymmetries, however, with $\hat{\overline{\kappa}}_\delta^c$ remaining flat across cohorts while the lower bound increases significantly, thereby providing weaker, but more robust evidence in favor of CSW's theme of increasing marital returns to higher education.

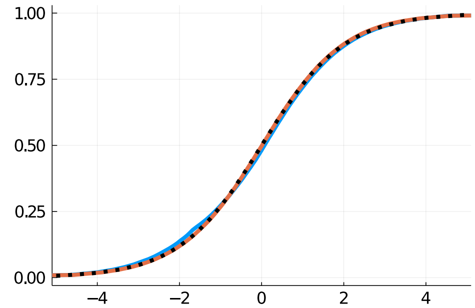
To interpret δ and understand better what is meant by "small" and "large" neighborhoods, Figure 2a plots the CDFs of $U_4 - U_0$ and $U_1 - U_3$ under the LFDs at which $\hat{\overline{\kappa}}_\delta^1$ is attained. The LFDs were computed as described in Section 4.3 using (25). Similar LFDs (not reported) were obtained for other U_a, U_b pairs, other cohorts, and the lower values $\hat{\underline{\kappa}}_\delta^c$. The CDFs of $U_4 - U_0$ and $U_1 - U_3$ under the LFDs are near indistinguishable from the logistic CDF (their distribution under F_*) when $\delta = 0.01$ and 0.10. With $\delta = 1$ the CDFs look close to logistic, while for $\delta = 10$ the CDF for $U_4 - U_0$ displays kinks and moves mass into the tails while the CDF for $U_1 - U_3$ is relatively less distorted.

To further aid the interpretation of δ, we compute the largest correlation between the elements of U under the LFDs for $\hat{\underline{\kappa}}_\delta^c$ and $\hat{\overline{\kappa}}_\delta^c$ as well as our *size* measure from Section 4.4. As these quantities are relatively stable across cohorts, we present their average across cohorts in Table 3. The shocks are uncorrelated under F_* and close to uncorrelated under the LFDs with $\delta = 0.01$ and 0.1, while for $\delta = 10$ some shocks (typically U_3 or U_4 and U_0) are strongly

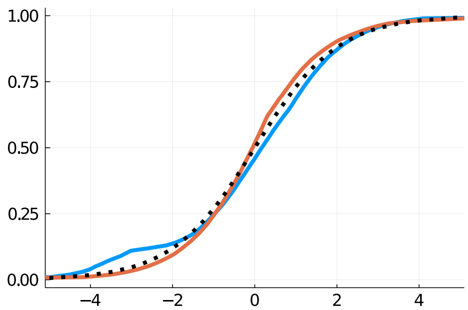
[21] One-sided 95% confidence sets are computed using our bootstrap procedure described in Section 6.2 with 1000 bootstrap replications for each cohort. We resample $\hat{P}_2^c$ using $\hat{P}_2^{c*} \sim N(\hat{P}_2^c, \hat{V}^c)$ where $\hat{V}^c$ is CSW's estimate of the covariance matrix of the sampling distribution of $\hat{P}_2^c$. Confidence sets constructed using Remark 6.3 were equivalent up to an order of magnitude of 10^{-3}.



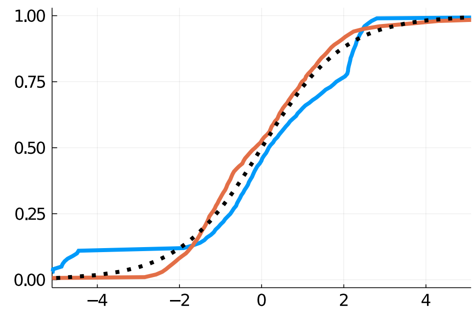
$\delta = 0.01.$



$\delta = 0.10.$

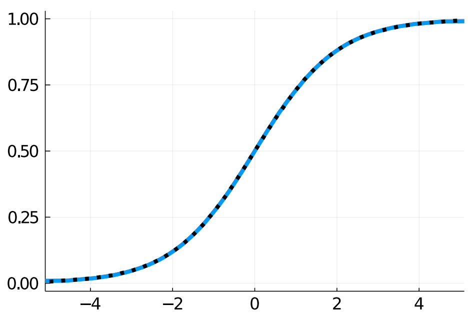


$\delta = 1.00.$

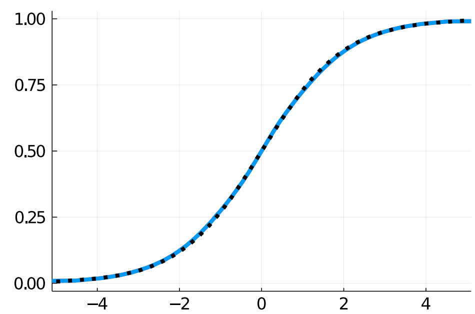


$\delta = 10.0.$

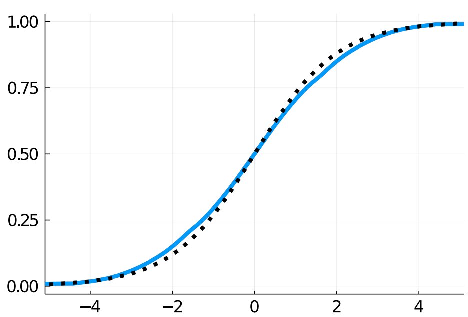
(a) First implementation (without exchangeability).



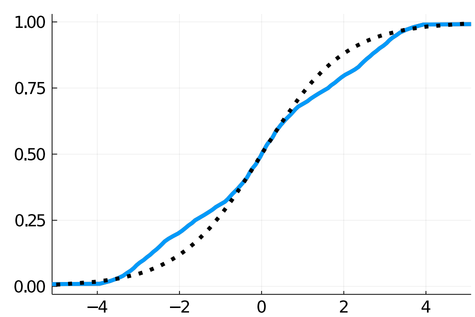
$\delta = 0.01.$



$\delta = 0.10.$



$\delta = 1.00.$



$\delta = 10.0.$

(b) Second implementation (dihedral exchangeability).

Figure 2: CDFs of $U_4 - U_0$ (blue solid lines) and $U_1 - U_3$ (orange solid lines) under the LFDs at which $\hat{\bar{\kappa}}^1_\delta$ is attained. The standard logistic CDF (black dotted lines) is also shown.

δ	First Implementation Max. corr., $\hat{\underline{\kappa}}_{\delta}^{c}$	Max. corr., $\hat{\overline{\kappa}}_{\delta}^{c}$	*size*	Second Implementation Max. corr., $\hat{\underline{\kappa}}_{\delta}^{c}$	Max. corr., $\hat{\overline{\kappa}}_{\delta}^{c}$	*size*
0.01	−0.016	−0.018	0.010	−0.018	0.016	0.006
0.10	−0.070	−0.072	0.037	−0.059	0.055	0.026
1.00	−0.246	−0.214	0.114	−0.138	0.093	0.099
10.0	−0.502	−0.493	0.240	−0.203	0.198	0.174

Table 3: Averages across cohorts of (i) the largest correlation of elements of U under the LFDs and (ii) our *size* measure from Section 4.4. Each is computed at the parameter values which minimize and maximize the sample criterion functions.

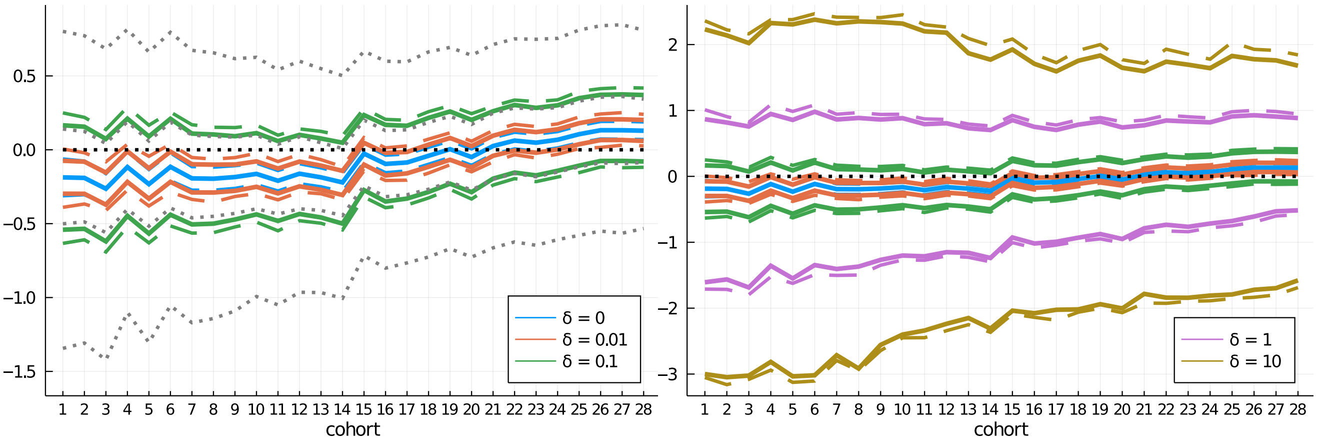
negatively correlated under the LFDs. Turning to the *size* measure, we see that the LFDs for $\delta = 0.01$ are distorting F_* in a manner that shifts the model-implied match probabilities by at most 0.010 (on average, across cohorts). By contrast, the LFDs for $\delta = 10$ are distorting F_* in a manner that shifts the match probabilities up to 0.24 (on average, across cohorts).

To investigate why the bounds become so wide under slight departures from F_*, Figure 6a plots estimates of κ^c when F^c is allowed to vary but θ^c is fixed at CSW's estimates under F_*, say $\hat{\theta}_*^c$. The bounds in Figure 6a for $\delta = 10$ are identical to those with $\delta = 100$ and represent the nonparametrically identified sets for κ^c when $\theta^c = \hat{\theta}_*^c$. These "fixed-$\theta$" sets are narrow compared with the bounds $[\underline{\kappa}_{\delta}^{c}, \overline{\kappa}_{\delta}^{c}]$ obtained when θ^c and F^c vary: they are roughly the same width as $[\underline{\kappa}_{0.01}^{c}, \overline{\kappa}_{0.01}^{c}]$. The cause of the wide bounds in Figure 1 is therefore the large variation in θ^c permitted when parametric assumptions on F^c are relaxed. As θ^c is just-identified from data on choice probabilities under any fixed distribution, overidentifying restrictions on θ^c and/or shape restrictions on F^c are required to further shrink the identified sets.

To this end, results for our second implementation are plotted in Figure 3. The dihedral exchangeability restriction shrinks the sets by around 67% for $\delta = 0.10$, 0.1, and 1, and around 53% for $\delta = 10$.[22] Estimates and confidence sets with $\delta = 0.01$ provide evidence that the SC to CG premium has indeed increased over time, corroborating CSW's findings under small departures from their i.i.d. Gumbel assumption. While bounds for larger neighborhoods are also very wide for this implementation, Figure 3b again shows a significant increase in the lower bound across cohorts.

CDFs of $U_4 - U_0$ and $U_1 - U_3$ under the LFDs for $\hat{\overline{\kappa}}^1$ for this implementation are plotted in Figure 2b. Here the LFDs for $U_4 - U_0$ and $U_1 - U_3$ are identical and symmetric due to the shape restriction. The CDFs of the LFDs with $\delta = 1$ or smaller are again virtually indistinguishable from the logistic CDF. For $\delta = 10$ the CDFs depart meaningfully from the logistic CDF but in a manner that is less kinked and irregular than Figure 2a, reflecting the fact that the shape restriction reduces asymmetries in the LFDs across types. Turning to Table 3, the *size* measure is around 3/4 the values for our first implementation because the shape restriction does not allow the parameters to vary as much when F_* is relaxed. The maximal correlations of the

[22]The fixed-θ sets are reduced by a similar amount; see Figure 6b.



(a) Small δ-neighborhoods. (b) Large δ-neighborhoods.

Figure 3: Sensitivity analysis for the SC to CG premium for white women for our second implementation. CSW's estimates under F_* are shown in blue ($\delta = 0$). Estimates of $\underline{\kappa}_\delta^c$ and $\overline{\kappa}_\delta^c$ are shown as solid lines; dashed lines are (pointwise) 95% confidence sets for $\underline{\kappa}_\delta^c$ and $\overline{\kappa}_\delta^c$. Grey dotted lines are estimates from our first implementation with $\delta = 0.01$ and $\delta = 0.10$.

shocks under the LFDs are also smaller, especially for larger δ, again reflecting the fact that the shape restriction reduces the asymmetries in the LFDs across types.[23]

We repeated the analysis for the "college-graduate" to "college-plus" premium and found qualitatively similar results (not reported). In particular, evidence for an increase in the premium across cohorts was found under small departures from F_* (i.e., $\delta = 0.01$) for the exchangeable implementation but not for the implementation without exchangeability. We also found a significant increase in the lower bounds on this premium across cohorts.

5.2 Welfare Analysis in a Rust Model

As a second empirical illustration, we perform a sensitivity analysis for counterfactual welfare in the DDC model of Rust (1987). This familiar setting serves as a useful laboratory in which to illustrate our procedure and some implementation issues that can arise in practice.

Model. The model is as described in Rust (1987). We focus on his specification in which maintenance costs are linear in the state (i.e., mileage). The counterfactual of interest is the change in average welfare arising from a 10% reduction in maintenance costs in each state.

Implementation. The model maps to our framework as described in Example 2.3. We focus on the implementation in Table IX of Rust (1987) in which the state space is 90 dimensional, $\beta = 0.9999$, and $\theta_\pi = (RC, MC)$ where RC is replacement cost and MC is the maintenance cost parameter (θ_{11} in Rust's notation). Payoffs are $\pi_{1,s}(\theta_\pi) = \tilde{\pi}_{1,s}(\theta_\pi) = -RC$ and

[23] Indeed, the correlation matrix under the LFD must be a symmetric circulant matrix due to the exchangeability restriction.

$\pi_{0,s}(\theta_\pi) = -0.001MC \times s$ and $\tilde{\pi}_{0,s}(\theta_\pi) = 0.9\pi_{0,s}(\theta_\pi)$. We normalize F so that the shocks have mean zero and the same variance as Rust's i.i.d. Gumbel specification (F_* hereafter) by appending the moment conditions $\mathbb{E}^F[U_j] = 0$ and $\mathbb{E}^F[U_j^2 - \pi^2/6] = 0$, for $j = 0, 1$ to g_4. The counterfactual function is $k(\theta, \gamma) = w'(\tilde{v} - v)$ where w is the stationary distribution of the state pre-intervention.

We estimate choice probabilities using Rust's Group 4 data. Nonparametric estimates of the 90 choice probabilities are noisy with zeros in many states (see also Figure 3 in Rust (1987)) whereas the model-implied CCPs are positive. We therefore proceed as described in Section 3.3, first estimating θ by maximum likelihood under F_* and then taking the model-implied CCPs at the MLE of θ as our estimate $\hat{P}$. As estimated CCPs are near zero for very low values of the state, we drop the moment conditions for CCPs in states where the CCP is less than 0.001 to avoid numerical instabilities induced by including near-degenerate moments. This reduces the dimension of g_2 by 19. In total there are 255 moments (71 for CCPs, 180 for Bellman equations, and 4 location/scale normalizations) and $\theta = (\theta_\pi, v, \tilde{v})$ has dimension 182.

We perform computations using our MPEC procedure described in Section 3.2. The inner optimization uses 75 moments (71 for CCPs, 4 location/scale normalizations) with the remaining 180 moments representing the Bellman equations for v and $\tilde{v}$ appended as constraints for the outer optimization. Expectations are computed numerically using $50,000$ scrambled Halton draws. Computation times are reported in Appendix C.2. As with the matching example, we define neighborhoods using hybrid of KL and χ^2 divergence (Assumption Φ(ii) fails for KL divergence for this example). We repeated our analysis with neighborhoods constrained by χ^2 and L^4 divergences as a robustness check. Overall, our findings are not sensitive to the choice of ϕ (see Appendix C.2 for a discussion).

Findings. Estimates and confidence sets for $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are plotted in Figure 4 for values of δ from 0.01 to 10.[24,25] As can be seen, the bounds expand rapidly under slight relaxations of the i.i.d. Gumbel assumption then stabilize by around $\delta = 1$. Indeed, with $\delta = 0.01$ the estimated bounds are $[44.6, 105.4]$ with confidence sets agreeing up to rounding error. The lower bound on change in average welfare is near zero, while the upper estimate for $\delta = 1$ is 160.5, approximately 220% the value under F_*.

To interpret δ, in Figure 5 we plot the CDFs of $U_1 - U_0$ under the LFDs at which $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained. The LFDs were computed as described in Section 4.3 using (26). The CDFs of $U_1 - U_0$ under the LFDs appear very close to logistic for $\delta = 0.01$ and therefore show that large differences in welfare counterfactuals can arise under very slight departures from

[24] Bounds with $\delta = 100$ were almost identical, so we truncate the figure at $\delta = 10$.

[25] One-sided 95% confidence sets are computed using our bootstrap procedure described in Section 6.2 with 1000 bootstrap replications. To resample CCPs, we draw $\theta_\pi^* \sim N(\hat{\theta}_\pi, \hat{V})$ where $\hat{\theta}_\pi$ is the MLE of (RC, MC) under F_* and $\hat{V}$ is the inverse Hessian, then take $\hat{P}_2^*$ as the model-implied CCPs under F_* at θ_π^*. Note that these confidence sets treat M_0, M_1, and w as deterministic. Confidence sets constructed using Remark 6.3 were equivalent up to an order of magnitude of 10^{-3}.

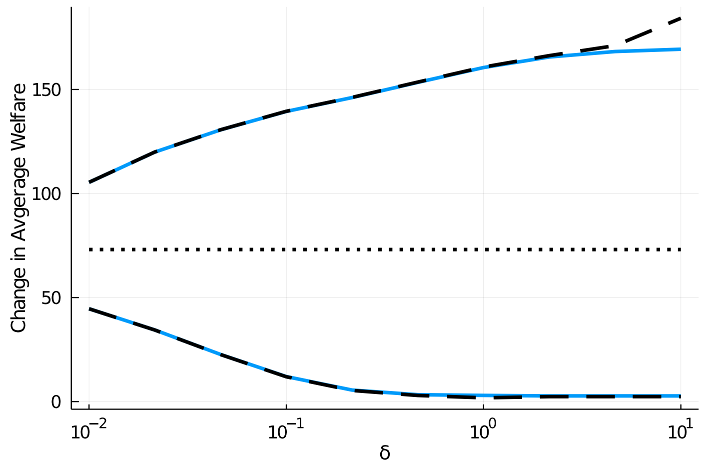


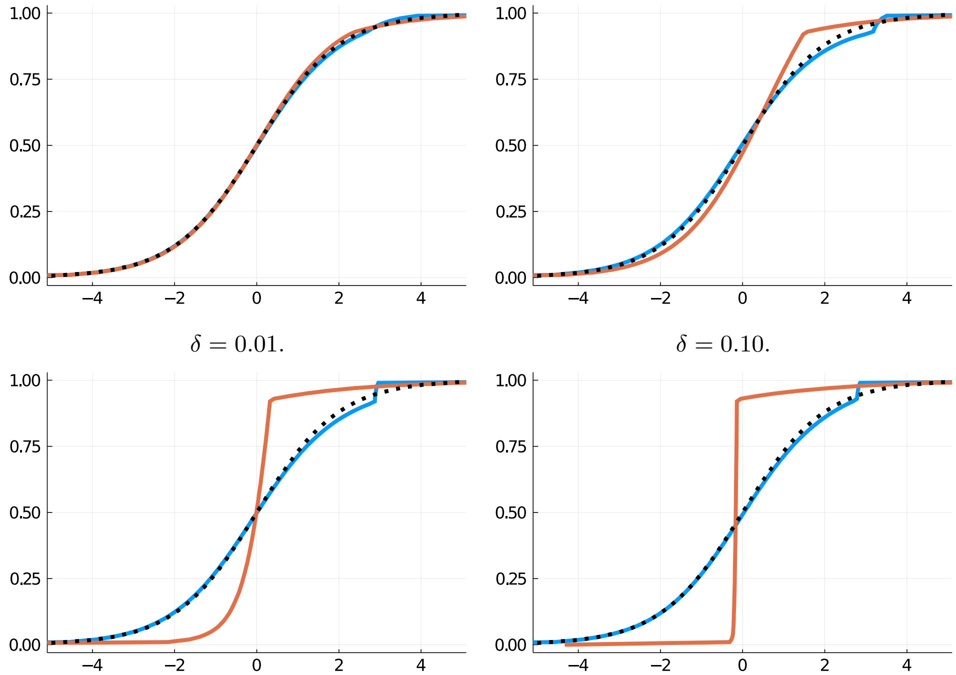
Figure 4: Sensitivity analysis for change in average welfare under a 10% maintenance cost subsidy. Estimates of $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are shown as solid lines; dashed lines are (pointwise) 95% confidence sets for $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$. Dotted line is the estimate under F_*.

	$\hat{\underline{\kappa}}_\delta$				$\hat{\overline{\kappa}}_\delta$			
δ	Corr(U_0,U_1)	*size*	*RC*	*MC*	Corr(U_0,U_1)	*size*	*RC*	*MC*
0.00	0.000	0.000	10.208	2.294	0.000	0.000	10.208	2.294
0.01	0.035	0.010	7.339	1.403	-0.032	0.014	13.401	3.306
0.10	-0.016	0.033	4.430	0.378	0.149	0.109	16.134	4.374
1.00	0.037	0.044	3.188	0.094	0.616	0.346	17.166	5.038
10.0	0.071	0.049	3.069	0.086	0.765	0.461	17.595	5.331

Table 4: Correlation of U_0 and U_1 under the LFD at which $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained, our *size* measure from Section 4.4, and replacement and maintenance cost parameters at which $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained.

the i.i.d. Gumbel assumption. The LFDs for $\hat{\overline{\kappa}}_\delta$ with large δ shift increasing amounts of mass to the center of the distribution of $U_1 - U_0$. If these LFDs seem too "rough" then smoothness restrictions could be imposed as described in Section 2, though any narrowing of the bounds would be due to the smoothness restriction rather than the structure of the model. The LFDs for $\hat{\underline{\kappa}}_\delta$ are relatively less distorted and are similar for $\delta = 0.1$, 1, and 10 because $\hat{\underline{\kappa}}_\delta$ stabilizes for smaller values of δ than $\hat{\overline{\kappa}}_\delta$ (cf. Figure 4).

Table 4 lists other metrics to help interpret the neighborhood size. The first is the correlation of U_0 and U_1 under the LFDs at which $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained. These are very small for $\delta = 0.01$ and remain small under the LFDs for $\hat{\underline{\kappa}}_\delta$ as δ increases, while the correlations under the LFDs for $\hat{\overline{\kappa}}_\delta$ increase dramatically. Given the asymmetry in distortions between the lower and upper values, we compute our *size* measure separately for both, using only the moments corresponding to the CCPs as these are most directly interpretable within the context of the model. We see that the LFDs for $\delta = 0.01$ are distorting F_* in a manner that shifts the model-implied CCPs by at most 0.014. By contrast, the LFDs for $\delta = 10$ distort F_* in a manner that



$\delta = 0.01.$ $\delta = 0.10.$

$\delta = 1.00.$ $\delta = 10.0.$

Figure 5: CDFs of $U_1 - U_0$ under the LFDs at which $\underline{\hat{\kappa}}_\delta$ (blue solid lines) and $\hat{\overline{\kappa}}_\delta$ (orange solid lines) are attained. The standard logistic CDF (black dotted lines) is also shown.

shifts model-implied CCPs by at most 0.05 for $\underline{\hat{\kappa}}_\delta$ and 0.46 for $\hat{\overline{\kappa}}_\delta$. Once again, this asymmetry reflects the fact that $\underline{\hat{\kappa}}_\delta$ stabilizes for smaller values of δ than $\hat{\overline{\kappa}}_\delta$.

The parameters at which $\underline{\hat{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained are also revealing. Table 4 presents MLEs of MC and RC, which are similar to the values reported in Table IX of Rust (1987). We see from Table 4 that $\underline{\hat{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained at very different parameter values than under F_*, with much smaller cost parameters for the lower bound and larger parameters for the upper bound, even for $\delta = 0.01$. Intuitively, a smaller MC means that the change in average welfare from the subsidy—which is proportional—must be small, and a low RC is needed to help the model to fit the pre-intervention CCPs at the smaller MC. While it is known that payoff parameters are not identified without parametric assumptions on F, it is perhaps surprising to see these parameters vary by so much under slight relaxations of F_*.

6 Estimation and Inference

In this section, we first establish consistency and the asymptotic distribution of the plug-in estimators $\underline{\hat{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ from Section 2.5. We then present two procedures for inference.

6.1 Large-sample Properties of Plug-in Estimators

We first introduce some regularity conditions to establish consistency of the plug-in estimators $\underline{\hat{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$. Recall the space $\mathcal{E}$ from Assumption Φ. We equip $\mathcal{E}$ with the Orlicz norm (see Appendix D)

$$\|f\|_\psi = \inf_{c>0} \frac{1}{c}\left(1 + \mathbb{E}^{F_*}[\psi(c|f(Z)|)]\right) .$$

The norm $\|\cdot\|_\psi$ is equivalent to the $L^2(F_*)$ norm for χ^2 and hybrid neighborhoods and the $L^q(F_*)$ norm for L^p neighborhoods ($p^{-1}+q^{-1}=1$), while for KL neighborhoods it is stronger than any $L^p(F_*)$ norm with $p<\infty$ but weaker than the sup-norm. We say that a class of functions $\{f_a : a \in \mathcal{A}\} \subset \mathcal{E}$ indexed by a metric space $\mathcal{A}$ is *$\mathcal{E}$-continuous in a* if $a' \to a$ in $\mathcal{A}$ implies $\|f_a - f_{a'}\|_\psi \to 0$. We shall also use a slightly stronger notion of constraint qualification than that which was introduced in Section 2.4:

Definition 6.1 Condition S' *holds at* (θ,γ,P) *if* $\vec{P} \in \mathrm{int}(\mathcal{G}(\theta,\gamma)+\mathcal{C})$.

Finally, let $\Theta_\delta(\gamma,P) = \{\theta \in \Theta : \Delta^\star(\theta;\gamma,P) < \delta\}$ where the program $\Delta^\star$ is defined in (16).

Assumption M (i) $k(\cdot;\theta,\gamma)$ and each entry of $g(\cdot;\theta,\gamma)$ are $\mathcal{E}$-continuous in (θ,γ);

(ii) $\mathbb{E}^{F_*}[\phi^\star(a_1 + a_2 k(U,\theta,\gamma) + a_3' g(U,\theta,\gamma))]$ is continuous in (θ,γ) for each $(a_1,a_2,a_3) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^d$;

(iii) $\Theta_\delta(\gamma_0,P_0)$ is nonempty and Condition S' holds at (θ,γ_0,P_0) for each $\theta \in \Theta_\delta(\gamma_0,P_0)$;

(iv) $\mathrm{cl}(\Theta_\delta(\gamma_0,P_0)) \supseteq \{\theta \in \Theta : \Delta^\star(\theta;\gamma_0,P_0) \le \delta\}$;

(v) Θ is a compact subset of $\mathbb{R}^{d_\theta}$.

Assumption M(i)(ii) are continuity conditions. If k and g consist of indicator functions of events, these conditions hold provided the probabilities of the events under F_* are continuous in (θ,γ). These conditions just require continuity in θ for models without γ. Nonemptyness of $\Theta_\delta(\gamma_0,P_0)$ is trivially satisfied when the model is correctly specified under F_*, i.e., there exists a $\theta\in\Theta$ solving (1) under F_*. Assumption M(iv) is made for convenience and can be relaxed; this condition simply ensures that there do not exist values of θ at which $\Delta^\star(\theta;\gamma_0,P_0)=\delta$ that are separated from $\Theta_\delta(\gamma_0,P_0)$. Assumption M(v) is standard.

Theorem 6.1 *Let Assumptions Φ and M hold and let $(\hat\gamma,\hat P) \to_p (\gamma_0,P_0)$ or, if there is no auxiliary parameter, $\hat P \to_p P_0$. Then: $\underline{\hat{\kappa}}_\delta \to_p \underline{\kappa}_\delta$ and $\hat{\overline{\kappa}}_\delta \to_p \overline{\kappa}_\delta$.*

We now derive the asymptotic distribution of the estimators. To simplify presentation, we assume the auxiliary parameter γ_0 is known and suppress dependence of all quantities on γ for the remainder of this section. This entails no loss of generality for models without auxiliary parameters, such as Examples 2.1 and 2.2 and the application in Section 5.1. In DDC models this presumes the law of motion of the state is known and so the asymptotic distribution reflects only sampling uncertainty from the estimated CCPs, which is the case for confidence sets reported when laws of motion and other inputs are first estimated "offline". Extending our approach to accommodate sampling variation in $\hat{\gamma}$ in a tractable manner appears to require exploiting application-specific structure of the model. We therefore defer such extensions to future work.

Define

$$\underline{b}_\delta(P) = \inf_{\theta \in \Theta_\delta(P)} \underline{K}_\delta(\theta; P) \,, \qquad \overline{b}_\delta(P) = \sup_{\theta \in \Theta_\delta(P)} \overline{K}_\delta(\theta; P) \,. \tag{28}$$

In this notation, we have $\underline{\kappa}_\delta = \underline{b}_\delta(P_0)$ and $\overline{\kappa}_\delta = \overline{b}_\delta(P_0)$ (see Lemma E.9 for a formal statement) and $\hat{\underline{\kappa}}_\delta = \underline{b}_\delta(\hat{P})$ and $\hat{\overline{\kappa}}_\delta = \overline{b}_\delta(\hat{P})$. We shall therefore derive the asymptotic distribution of the estimators by showing that $\underline{b}_\delta(P)$ and $\overline{b}_\delta(P)$ are directionally differentiable functions of P and applying a suitable delta method.

A function $f : \mathbb{R}^{d_1+d_2} \to \mathbb{R}$ is (Hadamard) directionally differentiable at P_0 if there is a continuous map $df_{P_0}[\cdot] : \mathbb{R}^{d_1+d_2} \to \mathbb{R}$ such that

$$\lim_{n\to\infty} \frac{f(P_0 + t_n h_n) - f(P_0)}{t_n} = df_{P_0}[h]$$

for all sequences $t_n \downarrow 0$ and $h_n \to h$ (Shapiro, 1990, p. 480). If $df_{P_0}[h]$ is linear in h then f is (fully) differentiable at P_0. To describe the directional derivatives, define

$$\begin{aligned}
\underline{\Xi}_\delta(\theta; P) &= \operatorname{argsup}_{\eta \geq 0, \zeta \in \mathbb{R}, \lambda \in \Lambda} - \mathbb{E}^{F_*}\Big[(\eta\phi)^\star(-k(U,\theta) - \zeta - \lambda' g(U,\theta))\Big] - \eta\delta - \zeta - \lambda_{12}' P \,, \\
\overline{\Xi}_\delta(\theta; P) &= \operatorname{arginf}_{\eta \geq 0, \zeta \in \mathbb{R}, \lambda \in \Lambda} \mathbb{E}^{F_*}\Big[(\eta\phi)^\star(k(U,\theta) - \zeta - \lambda' g(U,\theta))\Big] + \eta\delta + \zeta + \lambda_{12}' P \,,
\end{aligned}$$

where $(\eta\phi)^\star$ denotes the convex conjugate of $x \mapsto \eta \times \phi(x)$. Let $\underline{\Lambda}_\delta(\theta; P)$ and $\overline{\Lambda}_\delta(\theta; P)$ denote the projections of $\underline{\Xi}_\delta(\theta; P)$ and $\overline{\Xi}_\delta(\theta; P)$ for the subvector λ_{12} of (η, ζ, λ).[26] Finally, let $\underline{\Theta}_\delta(P_0) = \arg\min_{\theta \in \Theta} \underline{K}_\delta(\theta; P_0)$ and $\overline{\Theta}_\delta(P_0) = \arg\max_{\theta \in \Theta} \overline{K}_\delta(\theta; P_0)$. The sets $\underline{\Theta}_\delta(P_0)$ and $\overline{\Theta}_\delta(P_0)$ are nonempty and compact under Assumptions Φ and M. We require two further conditions:

Assumption M (continued) (vi) $\underline{\Theta}_\delta(P_0) \subseteq \Theta_\delta(P_0)$ and $\overline{\Theta}_\delta(P_0) \subseteq \Theta_\delta(P_0)$;
(vii) $\theta \mapsto \underline{\Lambda}_\delta(\theta; P_0)$ and $\theta \mapsto \overline{\Lambda}_\delta(\theta; P_0)$ are lower hemicontinuous at each $\theta \in \underline{\Theta}_\delta(P_0)$ and $\theta \in \overline{\Theta}_\delta(P_0)$, respectively.

Theorem 6.2 *Let Assumptions Φ and M hold. Then: the functions $\underline{b}_\delta(\cdot)$ and $\overline{b}_\delta(\cdot)$ are directionally differentiable at P_0, with*

$$d\underline{b}_{\delta,P_0}[h] = \min_{\theta \in \underline{\Theta}_\delta(P_0)} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta; P_0)} -\underline{\lambda}_{12}' h \,, \qquad d\overline{b}_{\delta,P_0}[h] = \max_{\theta \in \overline{\Theta}_\delta(P_0)} \min_{\overline{\lambda}_{12} \in \overline{\Lambda}_\delta(\theta; P_0)} \overline{\lambda}_{12}' h \,.$$

[26]That is, $\underline{\Lambda}_\delta(\theta; P) = \{(\lambda_1, \lambda_2) : (\eta, \zeta, \lambda_1, \lambda_2, \lambda_3, \lambda_4) \in \underline{\Xi}_\delta(\theta; P)\}$ and similarly for $\overline{\Lambda}_\delta(\theta; P)$.

In addition, if $\sqrt{n}(\hat{P} - P_0) \to_d N(0, \Sigma)$*, then*

$$\sqrt{n} \left[\begin{array}{c} \hat{\overline{\kappa}}_\delta - \overline{\kappa}_\delta \\ \hat{\underline{\kappa}}_\delta - \underline{\kappa}_\delta \end{array} \right] \to_d \left[\begin{array}{c} d\underline{b}_{\delta,P_0}[Z] \\ d\overline{b}_{\delta,P_0}[Z] \end{array} \right] , \text{ where } Z \sim N(0, \Sigma) \, .$$

Theorem 6.2 derives the asymptotic distribution of plug-in estimators by extending a result of Shapiro (2008).[27] The distribution is nonstandard due to the fact that $\underline{b}_\delta(\cdot)$ and $\overline{b}_\delta(\cdot)$ are directionally, but possibly not fully, differentiable. The exception is when $\cup_{\theta \in \underline{\Theta}_\delta(P_0)} \underline{\Lambda}_\delta(\theta; P_0) = \{\underline{\lambda}_{12}\}$, in which case $\sqrt{n}(\hat{\underline{\kappa}}_\delta - \underline{\kappa}_\delta) \to_d N(0, \underline{\lambda}_{12}' \Sigma \underline{\lambda}_{12})$, and similarly for the upper bound.

6.2 Inference Procedure 1: Bootstrap

In view of directional differentiability established in Theorem 6.2, it is well known that the bootstrap will be inconsistent (Dümbgen, 1993). We therefore specialize the general approach of Fang and Santos (2019) for inference on directionally differentiable functions to the present setting. Define

$$\widehat{d\underline{b}}_{\delta,P_0}[h] = \inf_{\theta \in \hat{\underline{\Theta}}_{\delta,n}} \sup_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta; \hat{P})} -\underline{\lambda}_{12}' h \, , \qquad \widehat{d\overline{b}}_{\delta,P_0}[h] = \sup_{\theta \in \hat{\overline{\Theta}}_{\delta,n}} \inf_{\overline{\lambda}_{12} \in \overline{\Lambda}_\delta(\theta; \hat{P})} \overline{\lambda}_{12}' h \, ,$$

where

$$\begin{aligned} \hat{\underline{\Theta}}_{\delta,n} &= \{\theta \in \Theta_\delta(\hat{P}) : \underline{K}_\delta(\theta; \hat{P}) \le \hat{\underline{\kappa}}_\delta + \hat{\nu}\sqrt{\log n / n}\} \, , \text{and} \\ \hat{\overline{\Theta}}_{\delta,n} &= \{\theta \in \Theta_\delta(\hat{P}) : \overline{K}_\delta(\theta; \hat{P}) \ge \hat{\overline{\kappa}}_\delta - \hat{\nu}\sqrt{\log n / n}\} \, . \end{aligned}$$

The quantity $\hat{\nu}$ is a (possibly sample-dependent) positive scalar tuning parameter that satisfies $\hat{\nu} \to_p \nu$ for some $\nu > 0$.[28] Any such $\hat{\nu}$ results in a confidence set with asymptotically correct coverage. We give some practical guidance on choice of $\hat{\nu}$ in Remark 6.2 below.

Let $\hat{P}^*$ denote a bootstrapped version of $\hat{P}$. In practice any bootstrap can be used provided it satisfies some mild consistency conditions below. For instance, in the empirical application in Section 5.1 we simply draw $\hat{P}^* \sim N(\hat{P}, \hat{\Sigma}/n)$ where $\hat{\Sigma}$ is a consistent estimator of Σ. Let

$$\hat{\underline{c}}_\alpha = \ \alpha\text{-quantile of } \widehat{d\underline{b}}_{\delta,P_0}[\sqrt{n}(\hat{P}^* - \hat{P})] \, , \qquad \hat{\overline{c}}_\alpha = \ \alpha\text{-quantile of } \widehat{d\overline{b}}_{\delta,P_0}[\sqrt{n}(\hat{P}^* - \hat{P})] \, ,$$

where the quantiles are computed by resampling $\hat{P}^*$. Lower, upper, and two-sided $100(1-\alpha)\%$

[27] The extensions allow for non-compact domain, non-convexity/concavity in θ, and discontinuity (which can arise at $\eta = 0$) and unboundedness (which can arise as $\eta \downarrow 0$, $|\zeta| \to \infty$, or $\|\lambda\| \to \infty$) of the objective. In independent work, Galvao and Parker (2019) also establish directional differentiability for minimax problems without convexity-concavity, though they require that the objective function is bounded.

[28] The approach of Fang and Santos (2019) will typically require some choice of tuning parameter.

confidence sets (CSs) for $\underline{\kappa}_\delta$, $\overline{\kappa}_\delta$, and $[\underline{\kappa}_\delta, \overline{\kappa}_\delta]$ are, respectively:

$$CS_{\delta,L}^{1-\alpha} = \left[\hat{\underline{\kappa}}_\delta - \frac{\hat{\underline{c}}_{1-\alpha}}{\sqrt{n}}, +\infty\right), \qquad CS_{\delta,U}^{1-\alpha} = \left(-\infty, \hat{\overline{\kappa}}_\delta - \frac{\hat{\overline{c}}_\alpha}{\sqrt{n}}\right],$$
$$CS_\delta^{1-\alpha} = \left[\hat{\underline{\kappa}}_\delta - \frac{\hat{\underline{c}}_{1-\alpha/2}}{\sqrt{n}}, \hat{\overline{\kappa}}_\delta - \frac{\hat{\overline{c}}_{\alpha/2}}{\sqrt{n}}\right].$$

We require a slight strengthening of Assumption M(vii) to establish validity of the procedure:

Assumption M (continued) (vii') $(\theta, P) \mapsto \underline{\Lambda}_\delta(\theta; P)$ and $(\theta, P) \mapsto \overline{\Lambda}_\delta(\theta; P)$ are lower hemicontinuous at (θ, P_0) for each $\theta \in \underline{\Theta}_\delta(P_0)$ and $\theta \in \overline{\Theta}_\delta(P_0)$, respectively.

We also require (standard) consistency and measurability conditions for the bootstrap for $\hat{P}^*$; see Assumption 3 of Fang and Santos (2019). Finally, let $\underline{G}_\delta$ and $\overline{G}_\delta$ denote the cumulative distribution functions of $d\underline{b}_{\delta,P_0}[Z]$ and $d\overline{b}_{\delta,P_0}[Z]$, respectively, with $Z \sim N(0, \Sigma)$.

Theorem 6.3 *Let Assumptions Φ and M(i)–(vi)(vii') hold, $\sqrt{n}(\hat{P} - P_0) \to_d N(0, \Sigma)$, and the bootstrap for $\hat{P}^*$ satisfy Assumption 3 of Fang and Santos (2019). Then: the bootstrap procedure is consistent for the asymptotic distribution derived in Theorem 6.2. In addition, if $\underline{G}_\delta$ and $\overline{G}_\delta$ are continuous and increasing at their $\alpha/2$, α, $1-\alpha$, and $1-\alpha/2$ quantiles, then:*

$$\lim_{n\to\infty} \Pr(\underline{\kappa}_\delta \in CS_{\delta,L}^{1-\alpha}) = 1 - \alpha, \qquad \lim_{n\to\infty} \Pr(\overline{\kappa}_\delta \in CS_{\delta,U}^{1-\alpha}) = 1 - \alpha,$$
$$\liminf_{n\to\infty} \Pr([\underline{\kappa}_\delta, \overline{\kappa}_\delta] \subseteq CS_\delta^{1-\alpha}) \geq 1 - \alpha.$$

Remark 6.1 Note $d\underline{b}_{\delta,P_0}$ and $d\bar{b}_{\delta,P_0}$ are convex and concave, respectively, when $\underline{\Theta}_\delta(P_0)$ and $\overline{\Theta}_\delta(P_0)$ are singletons. If so, $\liminf_{n\to\infty} \Pr_n(\underline{\kappa}_\delta \in CS_{\delta,L}^{1-\alpha}) \geq 1-\alpha$ and similarly for $CS_{\delta,U}^{1-\alpha}$ and $CS_\delta^{1-\alpha}$ where "$\Pr_n$" denotes probabilities along contiguous perturbations with $\sqrt{n}(\hat{P} - P_0) \to_d N(\mu, \Sigma)$ for fixed μ (Fang and Santos, 2019; Hong and Li, 2018).

Remark 6.2 Implementing our bootstrap procedure requires choosing $\hat{\nu}$. In principle, any $\hat{\nu}$ that satisfies $\hat{\nu} \to_p \nu > 0$ results in a CS with asymptotically correct coverage. As is evident from the construction of $\widehat{d\underline{b}}_{\delta,P_0}[\cdot]$ and $\widehat{d\underline{b}}_{\delta,P_0}[\cdot]$, choosing a smaller value of $\hat{\nu}$ results in (weakly) wider CSs. In the applications in Section 5, we set $\hat{\nu}$ to be equal to the minimum diagonal element of the covariance matrix (under F_*) of the moments evaluated at $(\hat{\theta}, \hat{\gamma}, \hat{P})$ where $\hat{\theta}$ is an estimate of the (point-identified) parameter θ_* solving (1) under F_*. We chose this quantity as it is related to the convexity of the inner optimization problem for small δ. Quantitatively, this resulted in $\hat{\nu}$ between 0.001 and 0.01. In practice, we recommend setting $\hat{\nu}$ to be of a similarly small magnitude, then performing a sensitivity analysis to check that critical values aren't too dependent on $\hat{\nu}$.

Remark 6.3 Choosing $\hat{\nu} = 0$ and replacing $\hat{\underline{\Theta}}_{\delta,n}$ and $\hat{\overline{\Theta}}_{\delta,n}$ by $\{\hat{\underline{\theta}}_\delta\}$ and $\{\hat{\overline{\theta}}_\delta\}$ where $\hat{\underline{\theta}}_\delta$ and $\hat{\overline{\theta}}_\delta$ minimize and maximize the sample criterions is valid but possibly conservative.

6.3 Inference Procedure 2: Projection

This second approach is computationally very simple but potentially conservative.[29] Suppose we have random vectors $\hat{P}_{1,U}^{1-\alpha}$, $\hat{P}_{2,U}^{1-\alpha}$, and $\hat{P}_{2,L}^{1-\alpha}$ that form a $100(1-\alpha)\%$ rectangular CS for P_0 (see below for its construction):

$$\liminf_{n\to\infty} \Pr\left(P_{10} \leq \hat{P}_{1,U}^{1-\alpha}\,,\, \hat{P}_{2,L}^{1-\alpha} \leq P_{20} \leq \hat{P}_{2,U}^{1-\alpha}\right) \geq 1-\alpha\,, \tag{29}$$

where the inequalities should be understood to hold element-wise. We replace the sample moment conditions by the following inequalities constructed from the rectangular CS for P_0:

$$\mathbb{E}^F[g_1(U,\theta)] \leq \hat{P}_{1,U}^{1-\alpha}\,, \quad \mathbb{E}^F[g_2(U,\theta)] \leq \hat{P}_{2,U}^{1-\alpha}\,, \quad \mathbb{E}^F[-g_2(U,\theta)] \leq -\hat{P}_{2,L}^{1-\alpha}\,. \tag{30}$$

Define

$$\underline{\hat{K}}_{\delta,1-\alpha}(\theta) = \left[\begin{array}{l} \underline{K}_{\delta,cs}(\theta;\hat{P}_{1-\alpha}) \\ +\infty \end{array}\right. , \quad \hat{\overline{K}}_{\delta,1-\alpha}(\theta) = \left[\begin{array}{ll} \overline{K}_{\delta,cs}(\theta;\hat{P}_{1-\alpha}) & \text{if } \Delta^\star_{cs}(\theta;\hat{P}_{1-\alpha}) < \delta, \\ -\infty & \text{if } \Delta^\star_{cs}(\theta;\hat{P}_{1-\alpha}) \geq \delta, \end{array}\right.$$

and $\underline{K}_{\delta,cs}$, $\overline{K}_{\delta,cs}$, and $\Delta^\star_{cs}$ are "relaxed" versions of (13), (14), and (16) formed by replacing sample counterparts of (1a) and (1b) by the inequalities (30). In the relaxed criterions, the set Λ is replaced by $\Lambda_{cs} := \mathbb{R}_+^{d_1+2d_2+d_3} \times \mathbb{R}^{d_4}$ because the number of inequality restrictions is now $d_1+2d_2+d_3$, the vector of moment functions g is replaced by $g_{cs} := (g_1, g_2, -g_2, g_3, g_4)$, P is replaced by $\hat{P}_{1-\alpha} := (\hat{P}_{1,U}^{1-\alpha}, \hat{P}_{2,U}^{1-\alpha}, -\hat{P}_{2,L}^{1-\alpha})$, and λ_{12} denotes the first d_1+2d_2 elements of λ.

Critical values are computed by optimizing the relaxed criterions:

$$\underline{\hat{\kappa}}_{\delta,1-\alpha} = \inf_{\theta\in\Theta} \underline{\hat{K}}_{\delta,1-\alpha}(\theta)\,, \quad \hat{\overline{\kappa}}_{\delta,1-\alpha} = \sup_{\theta\in\Theta} \hat{\overline{K}}_{\delta,1-\alpha}(\theta)\,.$$

Lower, upper, and two-sided $100(1-\alpha)\%$ CSs for $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are then given by

$$CS_{\delta,L}^{1-\alpha} = \left[\underline{\hat{\kappa}}_{\delta,1-\alpha}, +\infty\right), \quad CS_{\delta,U}^{1-\alpha} = \left(-\infty, \hat{\overline{\kappa}}_{\delta,1-\alpha}\right], \quad CS_{\delta}^{1-\alpha} = \left[\underline{\hat{\kappa}}_{\delta,1-\alpha/2}, \hat{\overline{\kappa}}_{\delta,1-\alpha/2}\right].$$

Note that $CS_\delta^{1-\alpha}$ requires projecting a $1-\alpha/2$ CS constructed from $\hat{P}_{1-\alpha/2}$.

Theorem 6.4 *Let Assumptions Φ and M(i),(iii)–(v) hold and $\hat{P}_{1-\alpha}$ and $\hat{P}_{1-\alpha/2}$ satisfy (29) with levels $1-\alpha$ and $1-\alpha/2$, respectively. Then:*

$$\liminf_{n\to\infty} \Pr(\underline{\kappa}_\delta \in CS_{\delta,L}^{1-\alpha}) \geq 1-\alpha\,, \qquad \liminf_{n\to\infty} \Pr(\overline{\kappa}_\delta \in CS_{\delta,U}^{1-\alpha}) \geq 1-\alpha\,,$$
$$\liminf_{n\to\infty} \Pr([\underline{\kappa}_\delta, \overline{\kappa}_\delta] \subseteq CS_\delta^{1-\alpha}) \geq 1-\alpha\,.$$

[29] We are grateful to a referee for suggesting this approach.

To construct a rectangular CS for P_0 satisfying (29), suppose $\sqrt{n}(\hat{P}-P_0) \to_d N(0,\Sigma)$ and we have a consistent estimator $\hat{\Sigma}$ of Σ. Let $\hat{\sigma}$ denote the vector formed by taking the square root of each diagonal entry of $\hat{\Sigma}$. Then, partition $\hat{\sigma}$ conformably as $\hat{\sigma} = (\hat{\sigma}_{(1)}, \hat{\sigma}_{(2)})$ and set

$$\hat{P}_{1,L}^{1-\alpha} = \hat{P}_1 + n^{-1/2}\hat{c}_{1-\alpha,1}\hat{\sigma}_{(1)}\,, \quad \hat{P}_{2,L}^{1-\alpha} = \hat{P}_2 - n^{-1/2}\hat{c}_{1-\alpha,2}\hat{\sigma}_{(2)}\,, \quad \hat{P}_{2,U}^{1-\alpha} = \hat{P}_2 + n^{-1/2}\hat{c}_{1-\alpha,2}\hat{\sigma}_{(2)}\,,$$

where the (scalar) critical values $\hat{c}_{1-\alpha,1}$ and $\hat{c}_{1-\alpha,2}$ solve

$$\Pr\left(\max_{1\le i\le d_1} Z_i/\hat{\sigma}_i \le \hat{c}_{1-\alpha,1}, \max_{d_1+1\le i\le d_2} |Z_i/\hat{\sigma}_i| \le \hat{c}_{1-\alpha,2}\right) = 1-\alpha\,, \quad Z \sim N(0,\hat{\Sigma})\,.$$

In particular, if $d_2 = 0$ then $\hat{c}_{1-\alpha,1}$ is the $(1-\alpha)$-quantile of $\max_{1\le i\le d_1} Z_i/\hat{\sigma}_i$ while if $d_1 = 0$ then $\hat{c}_{2,1-\alpha}$ is the $(1-\alpha)$-quantile of $\max_{1\le i\le d_2}|Z_i/\hat{\sigma}_i|$ with $Z \sim N(0,\hat{\Sigma})$.

7 Local Sensitivity

In this section, we introduce a measure of local sensitivity of the counterfactual κ with respect to F. We then contrast our approach with other recent approaches to local sensitivity.

7.1 Measure of Local Sensitivity

Our measure of local sensitivity of the counterfactual κ with respect to F at F_* is

$$s = \lim_{\delta\downarrow 0} \frac{(\overline{\kappa}_\delta - \underline{\kappa}_\delta)^2}{4\delta}\,.$$

If s is finite, then under the regularity conditions below, we have

$$\underline{\kappa}_\delta = \kappa_* - \sqrt{\delta s} + o(\sqrt{\delta})\,, \qquad\qquad \overline{\kappa}_\delta = \kappa_* + \sqrt{\delta s} + o(\sqrt{\delta})$$

as $\delta \downarrow 0$, where $\kappa_* = \mathbb{E}^{F_*}[k(U,\theta_*,\gamma_0)]$ and θ_* solves (1) under F_*.[30] We shall characterize s using an influence function representation and present an easy to compute estimator $\hat{s}$ of s.

To draw comparison with the local sensitivity literature, we restrict attention to moment equality models and impose (standard) regularity conditions. Assume that conditions (1b) and (1d) under F_* point identify a structural parameter $\theta_* \in \mathrm{int}(\Theta)$ and that Θ is compact. Note θ_* will depend on the parametric specification F_*. With some abuse of notation, let

$$g(u,\theta,\gamma,P_2) = \begin{bmatrix} g_2(u,\theta,\gamma) - P_2 \\ g_4(u,\theta,\gamma) \end{bmatrix}\,,$$

[30]Finiteness of s implies that κ_* is point identified. Note this may be true if θ_* is not point identified; see, e.g., Aguirregabiria (2005), Norets and Tang (2014), and Kalouptsidi et al. (2017) for DDC models.

$g_*(u) = g(u, \theta_*, \gamma_0, P_{20})$, and $k_*(u) = k(u, \theta_*, \gamma_0)$. Let $\mathbb{E}^{F_*}[g(U, \theta, \gamma_0, P_{20})]$ and $\mathbb{E}^{F_*}[k(U, \theta, \gamma_0)]$ be continuously differentiable with respect to θ at θ_*, $G := \frac{\partial}{\partial \theta'} \mathbb{E}^{F_*}[g(U, \theta, \gamma_0, P_{20})]\big|_{\theta=\theta_*}$ have full rank, $V := \mathbb{E}^{F_*}[g_*(U) g_*(U)']$ be finite and positive definite, $\mathbb{E}^{F_*}[k(U, \theta_*, \gamma_0)^2]$ be finite, and $k(\cdot, \theta, \gamma_0)$ and $g(\cdot, \theta, \gamma_0, P_{20})$ be $L^2(F_*)$ continuous in θ at θ_*.

We define the *influence function* of the counterfactual κ with respect to F at F_* as

$$\iota(u) = \mathbb{M} k_*(u) - J'(G'V^{-1}G)^{-1} G'V^{-1} g_*(u) \,, \tag{31}$$

where $\mathbb{M} k_*(u) := k_*(u) - \kappa_* - \mathbb{E}^{F_*}[k_*(U) g_*(U)'](V^{-1} - V^{-1}G(G'V^{-1}G)^{-1}G'V^{-1}) g_*(u)$ and $J := \frac{\partial}{\partial \theta} \mathbb{E}^{F_*}[k(U, \theta, \gamma_0)]\big|_{\theta=\theta_*}$. This notion is different from the statistical influence function of an estimator, as ι measures the sensitivity of an estimand to perturbations of a modeling assumption rather than the sensitivity of an estimator to perturbations of the data.[31]

The following theorem relates s and ι. We restrict attention to neighborhoods characterized by χ^2 divergence. Other ϕ-divergences are locally equivalent to χ^2 divergence, so this restriction entails no great loss of generality.[32]

Theorem 7.1 *Let the above GMM-type regularity conditions hold. Then:* $s = 2\mathbb{E}^{F_*}[\iota(U)^2]$.

In addition to reporting $\hat{\kappa} = \mathbb{E}^{F_*}[k(U, \hat{\theta}, \hat{\gamma})]$, the researcher could also report an estimate of the local sensitivity of the counterfactual with respect to F_*:

$$\hat{s} = 2\mathbb{E}^{F_*}[(\hat{k}(U) - \hat{\kappa})^2] + 2\hat{Q}'\hat{V}\hat{Q} - 4\mathbb{E}^{F_*}[\hat{g}(U)(\hat{k}(U) - \hat{\kappa})]'\hat{Q} \,,$$

where $\hat{k}(u) = k(u, \hat{\theta}, \hat{\gamma})$, $\hat{g}(u) = g(u, \hat{\theta}, \hat{\gamma}, \hat{P}_2)$, $\hat{V} = \mathbb{E}^{F_*}[\hat{g}(U)\hat{g}(U)']$, and

$$\hat{Q} = \mathbb{E}^{F_*}[\hat{k}(U)\hat{g}(U)'](\hat{V}^{-1} - \hat{V}^{-1}\hat{G}(\hat{G}'\hat{V}^{-1}\hat{G})^{-1}\hat{G}'\hat{V}^{-1}) + \hat{J}'(\hat{G}'\hat{V}^{-1}\hat{G})^{-1}\hat{G}'\hat{V}^{-1} \,,$$

with $\hat{G} = \frac{\partial}{\partial \theta'} \mathbb{E}^{F_*}[g(U, \theta, \hat{\gamma}, \hat{P}_2)]\Big|_{\theta=\hat{\theta}}$ and $\hat{J} = \frac{\partial}{\partial \theta} \mathbb{E}^{F_*}[k(U, \theta, \hat{\gamma})]\big|_{\theta=\hat{\theta}}$. Bounds on counterfactuals as F varies over small neighborhoods of F_* can then be estimated using $\hat{\kappa} \pm \sqrt{\delta \hat{s}}$. The estimator $\hat{s}$ is consistent under mild smoothness conditions:

Lemma 7.1 *Let the conditions of Theorem 7.1 hold. Also let* $(\hat{\theta}, \hat{\gamma}, \hat{P}_2) \to_p (\theta_*, \gamma_0, P_{20})$, *and let* $\mathbb{E}^{F_*}[g(U, \theta, \gamma, P_2) g(U, \theta, \gamma, P_2)']$, $\mathbb{E}^{F_*}[g(U, \theta, \gamma, P_2) k(U, \theta, \gamma)]$, $\mathbb{E}^{F_*}[g(U, \theta, \gamma, P_2)]$, $\mathbb{E}^{F_*}[k(U, \theta, \gamma)]$, $\frac{\partial}{\partial \theta'} \mathbb{E}^{F_*}[g(U, \theta, \gamma, P_2)]$, $\frac{\partial}{\partial \theta'} \mathbb{E}^{F_*}[k(U, \theta, \gamma)]$, *and* $\mathbb{E}^{F_*}[k(U, \theta, \gamma)^2]$ *each be continuous in* (θ, γ, P_2) *at* $(\theta_*, \gamma_0, P_{20})$. *Then:* $\hat{s} \to_p s$.

[31] We use the term influence function as $\kappa(F) - \kappa(F_*) \approx \int \iota \, \mathrm{d}(F - F_*)$ is valid for distributions F suitably close to F_*, where $\kappa(F)$ denotes the counterfactual under $U \sim F$. This mimics the asymptotic linear expansion for estimators, where F and F_* are replaced by the empirical and true probability measures, respectively.

[32] See Theorem 4.1 of Csiszár and Shields (2004). The quantity $2\mathbb{E}^{F_*}[\iota(U)^2]$ should be rescaled by a factor of $\phi''(1)$ for other ϕ divergences.

7.2 Comparison with Other Approaches

We conclude this section by comparing our approach with Andrews, Gentzkow, and Shapiro (2017, 2020; AGS hereafter) and Bonhomme and Weidner (2018; BW hereafter). In what follows, we shall restrict attention to models characterized only by moments of the form (1b) with $d_2 \geq d_\theta$ and in which the auxiliary parameter γ is vacuous. The approaches of AGS, BW, and this paper all apply to more general models; this restriction is simply to facilitate a comparison of their approaches with ours.

AGS consider a setting in which the moments (1b) are locally misspecified:

$$\mathbb{E}^{F_*}[g_2(U,\theta_*)] = P_{20} + n^{-1/2}c\,, \tag{32}$$

where c gives the direction of local misspecification. Suppose a researcher has a first-stage estimator $\hat{P}_2$ and computes an estimator $\hat{\theta}$ by minimizing

$$(\mathbb{E}^{F_*}[g_2(U,\theta)] - \hat{P}_2)'\hat{W}(\mathbb{E}^{F_*}[g_2(U,\theta)] - \hat{P}_2)$$

given some weight matrix $\hat{W} \to_p W$ where W is positive-definite and symmetric. The researcher would then estimate the counterfactual as $\hat{\kappa} = \mathbb{E}^{F_*}[k(U,\hat{\theta})]$, where $\hat{\theta}$ is the first-stage estimator. AGS's measure of *sensitivity* of $\hat{\kappa}$ to $\hat{P}_2$ is $J'(G'WG)^{-1}G'W$. The first-order asymptotic bias of $\hat{\kappa}$ due to local misspecification of the form (32) is therefore $J'(G'WG)^{-1}G'Wc$. AGS's measure of *informativeness* of $\hat{P}_2$ for $\hat{\kappa}$ is 1, meaning that all (statistical) variation in $\hat{\kappa}$ is explained by (statistical) variation in $\hat{P}_2$. Our measure of sensitivity instead characterizes the "specification variation" in κ as the researcher varies F while requiring that F satisfy (1b) for some $\theta \in \Theta$. AGS's measures and our measure of sensitivity therefore represent distinct but complementary quantities.

BW consider estimation of a target parameter (κ in our context) using a reference model $\mathcal{M}_R = \{(\theta,F) \in \Theta \times \{F_*\}\}$ when the true (θ_0, F_0) possibly belongs to a larger model $\mathcal{M}_L = \{(\theta,F) \in \Theta \times \mathcal{N}_\delta\}$ where $\delta \downarrow 0$ with $n\delta \to \tau$ for some constant $\tau > 0$. Shrinking $\delta \downarrow 0$ with n lends tractability to BW's approach but is not necessarily appropriate in settings such as ours where F is not nonparametrically identified. BW seek estimators of κ under $\mathcal{M}_R$ that minimize worst-case asymptotic bias or MSE over $\mathcal{M}_L$. Consider the one-step estimator

$$\hat{\kappa} = \mathbb{E}^{F_*}[k(U,\hat{\theta})] + a'(\mathbb{E}^{F_*}[g_2(U,\hat{\theta})] - \hat{P}_2)\,,$$

where $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ_* and $a \in \mathbb{R}^{d_2}$ satisfies $J' = -a'G$ so that $\hat{\kappa}$ does not depend asymptotically on $\hat{\theta}$. The true counterfactual is $\kappa_0 = \mathbb{E}^{F_0}[k(U,\theta_0)]$ where $(\theta_0, F_0) \in \mathcal{M}_L$ satisfies $\mathbb{E}^{F_0}[g_2(U,\theta_0)] = P_{20}$. If $\mathcal{M}_R$ is correctly specified so that $\mathbb{E}^{F_*}[g_2(U,\theta_*)] = P_{20}$, then for any a the worst-case asymptotic bias of the one-step estimator is

$$\lim_{n\to\infty} \sup_{(\theta_0,F_0)\in\mathcal{M}_L:\mathbb{E}^{F_0}[g_2(U,\theta_0)]=P_{20}} |\sqrt{n}(\kappa_* - \kappa_0))| = \sqrt{\tau s}\,,$$

where s is our measure of local sensitivity. If we allow for local misspecification of $\mathcal{M}_R$, in the sense that $\mathbb{E}^{F_*}[g_2(U,\theta_*)] \neq P_{20}$, then the worst-case asymptotic bias of the one-step estimator is

$$\lim_{n\to\infty} \sup_{(\theta_0,F_0)\in\mathcal{M}_L:\mathbb{E}^{F_0}[g_2(U,\theta_0)]=P_{20}} \left|\sqrt{n}\left(\kappa_* - \kappa_0 + a'\left(\mathbb{E}^{F_*}[g_2(U,\theta_*)] - P_{20}\right)\right)\right| = \sqrt{\tau s_a}\,,$$

where s_a is our local sensitivity measure with k replaced by $k + a'g_2$.

8 Conclusion

This paper introduces a tractable framework for performing a global sensitivity analysis of counterfactuals to researchers' assumptions about the distribution of latent variables in a class of structural models. In particular, we show how to construct the smallest and largest counterfactuals obtained as the distribution of unobservables varies over fully nonparametric neighborhoods of the researcher's parametric specification while other structural features of the model are maintained. Our procedure recovers sharp bounds on the nonparametrically identified set of counterfactuals as the neighborhoods size becomes large and has connections with local sensitivity analyses over small neighborhoods. We illustrate our procedure with empirical applications to matching models and dynamic discrete choice.

References

Aguirregabiria, V. (2005). Nonparametric identification of behavioral responses to counterfactual policy interventions in dynamic discrete decision processes. *Economics Letters 87*(3), 393–398.

Aguirregabiria, V. and P. Mira (2007). Sequential estimation of dynamic discrete games. *Econometrica 75*(1), 1–53.

Allen, R. and J. Rehbeck (2019). Identification with additively separable heterogeneity. *Econometrica 87*(3), 1021–1054.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017). Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics 132*(4), 1553–1592.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2020). On the informativeness of descriptive statistics for structural estimates. *Econometrica* (6), 2231–2258.

Armstrong, T. B. and M. Kolesár (2021). Sensitivity analysis using approximate moment condition models. *Quantitative Economics 12*(1), 77–108.

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

Beresteanu, A., I. Molchanov, and F. Molinari (2011). Sharp identification regions in models with convex moment predictions. *Econometrica 79*(6), 1785–1821.

Berger, J. O. (1984). The robust Bayesian viewpoint (with discussion). In J. B. Kadane (Ed.), *Robustness of Bayesian analyses*. North-Holland.

Berger, J. O. and L. M. Berliner (1986). Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. *The Annals of Statistics 14*(2), 461–486.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica 60*(4), 889–917.

Bonhomme, S. and M. Weidner (2018). Minimizing sensitivity to model misspecification. cemmap Working Paper, cwp59/18.

Bresnahan, T. F. and P. C. Reiss (1990). Entry in monopoly market. *The Review of Economic Studies 57*(4), 531–553.

Bresnahan, T. F. and P. C. Reiss (1991). Empirical models of discrete games. *Journal of Econometrics 48*(1), 57–81.

Chamberlain, G. and E. E. Leamer (1976). Matrix weighted averages and posterior bounds. *Journal of the Royal Statistical Society. Series B (Methodological) 38*(1), 73–84.

Chen, X., E. Tamer, and A. Torgovitsky (2011). Sensitivity analysis in semiparametric likelihood models. Cowles Foundation Discussion Paper no. 1836.

Chesher, A. and A. M. Rosen (2017). Generalized instrumental variable models. *Econometrica 85*(3), 959–989.

Chiappori, P.-A., B. Salanié, and Y. Weiss (2017). Partner choice, investment in children, and the marital college premium. *American Economic Review 107*(8), 2109–2167.

Chiong, K., Y.-W. Hsieh, and M. Shum (2017). Counterfactual evaluation in semiparametric multinomial choice models. USC-INET research paper no. 17-20.

Choo, E. and A. Siow (2006). Who marries whom and why. *Journal of Political Economy 114*(1), 175–201.

Ciliberto, F. and E. Tamer (2009). Market structure and multiple equilibria in airline markets. *Econometrica 77*(6), 1791–1828.

Csiszár, I. and P. C. Shields (2004). Information theory and statistics: A tutorial. *Communications and Information Theory 1*(4), 417–528.

Dagsvik, J. K. (2000). Aggregation in matching markets. *International Economic Review 41*(1), 27–58.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields 95*(1), 125–140.

Ekeland, I., A. Galichon, and M. Henry (2010). Optimal transportation and the falsifiability of incompletely specified economic models. *Economic Theory 42*(2), 355–374.

Fang, Z. and A. Santos (2019). Inference on directionally differentiable functions. *Review of Economic Studies 86*(1), 377–412.

Galichon, A. and M. Henry (2011). Set Identification in Models with Multiple Equilibria. *The Review of Economic Studies 78*(4), 1264–1298.

Galichon, A. and B. Salanié (2020). Cupid's invisible hand: Social surplus and identification in matching models. `http://dx.doi.org/10.2139/ssrn.1804623`.

Galvao, A. F. and T. Parker (2019). Uniform inference for bounds on the distribution and quantile functions of treatment effects in randomized experiments. arxiv:1911.10215 [econ.em].

Giacomini, R., T. Kitagawa, and H. Uhlig (2016). Estimation under ambiguity. Technical report, UCL and Chicago.

Grieco, P. L. E. (2014). Discrete games with flexible information structures: an application to local grocery markets. *The RAND Journal of Economics 45*(2), 303–340.

Hansen, L. P. and T. J. Sargent (2001). Robust control and model uncertainty. *The American Economic Review 91*(2), 60–66.

Hansen, L. P. and T. J. Sargent (2008). *Robustness*. Princeton.

Ho, P. (2018). Global robust bayesian analysis in large models. Technical report, Princeton.

Hong, H. and J. Li (2018). The numerical delta method. *Journal of Econometrics 206*(2), 379–394.

Kalouptsidi, M., Y. Kitamura, L. Lima, and E. Souza-Rodrigues (2020). Partial identification and inference for dynamic models and counterfactuals. *NBER Working Paper No. 26761*.

Kalouptsidi, M., P. T. Scott, and E. Souza-Rodrigues (2017). Identification of counterfactuals in dynamic discrete choice models. *NBER Working Paper No. 21527*.

Keane, M. P., P. E. Todd, and K. I. Wolpin (2011, May). *The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications*, Volume 4 of *Handbook of Labor Economics*, Chapter 4, pp. 331–461. Elsevier.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013). Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica 81*(3), 1185–1201.

Kline, B. and E. Tamer (2016). Bayesian inference in a class of partially identified models. *Quantitative Economics 7*(2), 329–366.

Komunjer, I. and G. Ragusa (2016). Existence and characterization of conditional density projections. *Econometric Theory 32*(4), 947–987.

Lafférs, L. (2019). Bounding average treatment effects using linear programming. *Empirical Economics 57*(3), 727–767.

Leamer, E. E. (1982). Sets of posterior means with bounded variance priors. *Econometrica 50*(3), 725–736.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review 75*(3), 308–313.

Lemieux, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer, New York.

Li, L. (2018). Identification of structural and counterfactual parameters in a large class of structural econometric models. Technical report, Pennsylvania State University.

Liese, F. and I. Vajda (1987). *Convex statistical distances*. Teubner, Leipzig.

Maccheroni, F., M. Marinacci, and A. Rustichini (2006). Ambiguity aversion, robustness, and the variational representation of preferences. *Econometrica 74*(6), 1447–1498.

Manski, C. F. (2007). Partial identification of counterfactual choice probabilities. *International Economic Review 48*(4), 1393–1410.

Manski, C. F. (2014). Identification of income–leisure preferences and evaluation of income tax policy. *Quantitative Economics 5*(1), 145–174.

McFadden, D. L. (1974). *Conditional Logit Analysis of Qualitative Choice Behavior*, pp. 105–142. Frontiers in Econometrics. Academic Press.

McFadden, D. L. (1978). *Modelling the Choice of Residential Location*, pp. 75–96. Spatial Interaction Theory and Planning Models. North Holland: Amsterdam.

Mukhin, Y. (2018). Sensitivity of regular estimators. arxiv:1805.08883 [econ.em].

Norets, A. and X. Tang (2014). Semiparametric inference in dynamic binary choice models. *The Review of Economic Studies 81*(3), 1229–1262.

Owen, A. B. (2017). A randomized Halton algorithm in R. arxiv:1706.02808 [stat.co].

Qiao, Y. and N. Minematsu (2010). A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing 58*(7), 3884–3890.

Rust, J. (1987). Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica 55*(5), 999–1033.

Schennach, S. M. (2014). Entropic latent variable integration via simulation. *Econometrica 82*(1), 345–385.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications 66*(3), 477–487.

Shapiro, A. (2008). Asymptotics of minimax stochastic programs. *Statistics & Probability Letters 78*(2), 150–157.

Su, C.-L. and K. L. Judd (2012). Constrained optimization approaches to estimation of structural models. *Econometrica 80*(5), 2213–2230.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies 70*(1), 147–165.

Tebaldi, P., A. Torgovitsky, and H. Yang (2019). Nonparametric estimates of demand in the california health insurance exchange. NBER working paper.

Torgovitsky, A. (2019a). Nonparametric inference on state dependence in unemployment. *Econometrica 87*(5), 1475–1505.

Torgovitsky, A. (2019b). Partial identification by extending subdistributions. *Quantitative Economics 10*(1), 105–144.

Online Appendix to "Counterfactual Sensitivity and Robustness"

Timothy Christensen Benjamin Connault

A Extensions

In this section we present three extensions of the methodology developed in the main text. The first two extensions are to different classes of model, namely conditional moment models and nonseparable models. The final extension shows how to impose exchangeability as a shape restriction on the class of distributions.

A.1 Conditional Moments

Consider the conditional moment model

$$\begin{array}{ll} \mathbb{E}^F[g_1(U,X,\theta,\gamma_0)|X=x] \leq P_{10,x}, & \mathbb{E}^F[g_2(U,X,\theta,\gamma_0)|X=x] = P_{20,x}, \\ \mathbb{E}^F[g_3(U,X,\theta,\gamma_0)|X=x] \leq 0, & \mathbb{E}^F[g_4(U,X,\theta,\gamma_0)|X=x] = 0, \end{array} \quad \text{for each } x \in \mathcal{X} \tag{33}$$

where the conditioning variable X takes values in a finite set $\mathcal{X}$. Suppose we are interested in a counterfactual[33]

$$\kappa = \sum_{x \in \mathcal{X}} \mathbb{E}^F[k(U,X,\theta,\gamma_0)|X=x]\,. \tag{34}$$

Suppose a researcher assumes $U|X = x \sim F_*$ for each x. We show how to bound counterfactuals when we relax this parametric assumption and allow each conditional distribution of U given $X = x$, say F_x, to depart from F_* and vary in a neighborhood $\mathcal{N}_{\delta_x}$ of F_*. Thus, in relaxing the parametric assumption F_* we are allowing the conditional distributions F_x to vary with x, and therefore relaxing independence of U and X.[34] For simplicity we assume each $\mathcal{N}_\delta$ is defined by the same ϕ function, though we allow the neighborhood size to possibly vary with x.

Let $\boldsymbol{\delta} = (\delta_x)_{x\in\mathcal{X}}$. By analogy with (8) and (9), we are interested in the values

$$\underline{\kappa}_{\boldsymbol{\delta}} = \inf_{\theta\in\Theta,(F_x\in\mathcal{N}_{\delta_x})_{x\in\mathcal{X}}} \sum_{x\in\mathcal{X}} \mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{subject to (33)}, \tag{35}$$

$$\overline{\kappa}_{\boldsymbol{\delta}} = \sup_{\theta\in\Theta,(F_x\in\mathcal{N}_{\delta_x})_{x\in\mathcal{X}}} \sum_{x\in\mathcal{X}} \mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{subject to (33)}. \tag{36}$$

[33]Note κ can be the expected value at a particular value x_0 if $k(U,x,\theta,\gamma_0) = 0$ for $x \neq x_0$. More generally, κ can be a weighted average by incorporating the weighting into the definition of $k(u,x,\theta,\gamma_0)$.

[34]The case with U independent of X is subsumed in (1) by stacking the moment functions and reduced-form parameters by values of the conditioning variable: $g_1(U,\theta,\gamma) = (g_1(U,x,\theta,\gamma))_{x\in\mathcal{X}}$ and similarly for g_2, g_3, and g_4, $P_{10} = (P_{10,x})_{x\in\mathcal{X}}$, $P_{20} = (P_{20,x})_{x\in\mathcal{X}}$, and $k(U,\theta,\gamma) = \sum_{x\in\mathcal{X}} k(U,x,\theta,\gamma)$.

By similar reasoning to Section 2.3, the optimization problems defining $\underline{\kappa}_{\boldsymbol{\delta}}$ and $\overline{\kappa}_{\boldsymbol{\delta}}$ can be written as optimizations of profiled criterion functions $\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma_0,P_0)$ and $\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma_0,P_0)$ with respect to θ. Let $P=(P_x)_{x\in\mathcal{X}}$ where each $P_x=(P_{1,x},P_{2,x})$ is partitioned conformably with g_1 and g_2. For a generic (θ,γ,P), the profiled criterion functions are

$$\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P)=\inf_{(F_x\in\mathcal{N}_{\delta_x})_{x\in\mathcal{X}}}\sum_{x\in\mathcal{X}}\mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{s.t. (33) holding at } (\theta,\gamma,P)\,,$$
$$\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P)=\sup_{(F_x\in\mathcal{N}_{\delta_x})_{x\in\mathcal{X}}}\sum_{x\in\mathcal{X}}\mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{s.t. (33) holding at } (\theta,\gamma,P)\,.$$

These profiled criterion functions have a dual formulation analogous to Proposition 2.1. Let $g(\cdot,x,\theta,\gamma)=(g_1(\cdot,x,\theta,\gamma),\ldots,g_4(\cdot,x,\theta,\gamma))$ denote the vector of moment functions evaluated at $X=x$. Recall that $d=\sum_{i=1}^4 d_i$ where d_i is the dimension of g_i, $i=1,\ldots,4$, and that $\Lambda=\mathbb{R}_+^{d_1}\times\mathbb{R}^{d_2}\times\mathbb{R}_+^{d_3}\times\mathbb{R}^{d_4}$, and let $\lambda_{12,x}$ denote the first d_1+d_2 elements of $\lambda_x\in\Lambda$.

Assumption $\boldsymbol{\Phi}$-conditional (i) $\phi\in\Phi_0$.

(ii) $k(\,\cdot\,,x,\theta,\gamma)$ and each entry of $g(\,\cdot\,,x,\theta,\gamma)$ belong to $\mathcal{E}$ for each $\theta\in\Theta$, $\gamma\in\Gamma$, and $x\in\mathcal{X}$.

Proposition A.1 *Let Assumption Φ-conditional hold. If there exist distributions $F_x\in\mathcal{N}_{\delta_x}$ for each $x\in\mathcal{X}$ such that moment conditions (33) hold at (θ,γ,P), then*

$$\begin{aligned}&\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P) &(37)\\ &=\sup_{(\eta_x>0,\zeta_x\in\mathbb{R},\lambda_x\in\Lambda)_{x\in\mathcal{X}}}\sum_{x\in\mathcal{X}}\left(-\eta_x\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,x,\theta,\gamma)+\zeta_x+\lambda_x'g(U,x,\theta,\gamma)}{-\eta_x}\right)\right]-\eta_x\delta_x-\zeta_x-\lambda_{12,x}'P_x\right)\,,\\ &\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P) &(38)\\ &=\inf_{(\eta_x>0,\zeta_x\in\mathbb{R},\lambda_x\in\Lambda)_{x\in\mathcal{X}}}\sum_{x\in\mathcal{X}}\left(\eta_x\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,x,\theta,\gamma)-\zeta_x-\lambda_x'g(U,x,\theta,\gamma)}{\eta_x}\right)\right]+\eta_x\delta_x+\zeta_x+\lambda_{12,x}'P_x\right)\,.\end{aligned}$$

Moreover, if the value of problem (37) is $+\infty$ (equivalently, if the value of problem (38) is $-\infty$), then for at least one $x\in\mathcal{X}$ there is no distribution in $\mathcal{N}_{\delta_x}$ under which the moment conditions (33) hold at (θ,γ,P).

Analogously to Section 2.5, estimators $\hat{\underline{\kappa}}_{\boldsymbol{\delta}}$ and $\hat{\overline{\kappa}}_{\boldsymbol{\delta}}$ of $\underline{\kappa}_{\boldsymbol{\delta}}$ and $\overline{\kappa}_{\boldsymbol{\delta}}$ are computed by minimizing and maximizing sample criterions with respect to θ. Letting $\hat{P}=(\hat{P}_x)_{x\in\mathcal{X}}$, the sample criterions are

$$\hat{\underline{K}}_{\boldsymbol{\delta}}(\theta)=\left[\begin{array}{l}\underline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P)\\ +\infty\end{array}\right.,\quad \hat{\overline{K}}_{\boldsymbol{\delta}}(\theta)=\left[\begin{array}{ll}\overline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P) & \text{if } \Delta_x^\star(\theta;\hat\gamma,\hat P_x)<\delta_x \text{ for each } x\in\mathcal{X},\\ -\infty & \text{if } \Delta_x^\star(\theta;\hat\gamma,\hat P_x)\geq\delta_x \text{ for some } x\in\mathcal{X},\end{array}\right.$$

where $\overline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P)$ and $\overline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P)$ denote the dual representations from Proposition A.1 evaluated at $(\hat{\gamma}, \hat{P})$, and $\Delta_x^\star(\theta; \hat{\gamma}, \hat{P}_x)$ denotes the program

$$\Delta_x^\star(\theta; \gamma, P_x) = \sup_{\zeta_x \in \mathbb{R}, \lambda_x \in \Lambda} -\mathbb{E}^{F_*}\left[\phi^\star(-\zeta_x - \lambda_x' g(U, x, \theta, \gamma))\right] - \zeta_x - \lambda_{12,x}' P_x$$

evaluated at $(\hat{\gamma}, \hat{P}_x)$. Consistency of the estimators and an asymptotic distribution theory may be derived by a suitable modification of the arguments in Section 6.

A.2 Nonseparable Moments

Here we show that our approach also extends to nonseparable models of the form

$$\begin{aligned} &\mathbb{E}^H[\tilde{g}_1(U, X, \theta, \tilde{\gamma}_0)] \le P_{10}, \quad && \mathbb{E}^H[\tilde{g}_2(U, X, \theta, \tilde{\gamma}_0)] = P_{20}, \\ &\mathbb{E}^H[\tilde{g}_3(U, X, \theta, \tilde{\gamma}_0)] \le 0, && \mathbb{E}^H[\tilde{g}_4(U, X, \theta, \tilde{\gamma}_0)] = 0, \end{aligned} \tag{39}$$

where the expectation is with respect to the joint distribution H of (U, X) and X again takes values in a finite set $\mathcal{X}$. We also consider a counterfactual function

$$\kappa = \mathbb{E}^H[\tilde{k}(U, X, \theta, \tilde{\gamma}_0)]\,. \tag{40}$$

As in the previous subsection, we suppose the researcher assumes $U|X = x \sim F_*$ for each x as this independent specification is often used in applied work. We wish to relax the researcher's parametric assumption and allow $U|X = x \sim F_x$ where each F_x is in a neighborhood $\mathcal{N}_{\delta_x}$ of F_*. For simplicity we shall again assume each $\mathcal{N}_\delta$ is defined by the same ϕ function, though we allow the neighborhood size to possibly vary with x.

Write $H(u, x) = q_{0,x} \times F_x(u)$ where $q_{0,x} = \Pr(X = x)$. The probabilities $q_0 = (q_{0,x})_{x \in \mathcal{X}}$ can be consistently estimated from data on X. Let $\gamma_0 = (\tilde{\gamma}_0, q_0)$ and $g_1(U, x, \theta, \gamma_0) = q_{0,x} \times \tilde{g}_1(U, x, \theta, \tilde{\gamma}_0)$ and similarly for g_2, g_3, g_4, and k. The model (39) and counterfactual (40) can then be written in similar notation to the previous subsection, namely as

$$\begin{aligned} &\sum_x \mathbb{E}^{F_x}[g_1(U, x, \theta, \gamma_0)] \le P_{10}, \quad && \sum_x \mathbb{E}^{F_x}[g_2(U, x, \theta, \gamma_0)] = P_{20}, \\ &\sum_x \mathbb{E}^{F_x}[g_3(U, x, \theta, \gamma_0)] \le 0, && \sum_x \mathbb{E}^{F_x}[g_4(U, x, \theta, \gamma_0)] = 0, \end{aligned} \tag{41}$$

and $\kappa = \sum_x \mathbb{E}^{F_x}[k(U, x, \theta, \gamma_0)]$. With $\boldsymbol{\delta} = (\delta_x)_{x \in \mathcal{X}}$, our objects of interest are then

$$\begin{aligned} \underline{\kappa}_{\boldsymbol{\delta}} &= \inf_{\theta \in \Theta, (F_x \in \mathcal{N}_{\delta_x})_{x \in \mathcal{X}}} \sum_x \mathbb{E}^{F_x}[k(U, x, \theta, \gamma_0)] \quad \text{subject to (41)}, \\ \overline{\kappa}_{\boldsymbol{\delta}} &= \sup_{\theta \in \Theta, (F_x \in \mathcal{N}_{\delta_x})_{x \in \mathcal{X}}} \sum_x \mathbb{E}^{F_x}[k(U, x, \theta, \gamma_0)] \quad \text{subject to (41)}. \end{aligned}$$

As before, the optimization problems defining $\underline{\kappa}_{\boldsymbol{\delta}}$ and $\overline{\kappa}_{\boldsymbol{\delta}}$ can be written as optimizations of profiled criterion functions $\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma_0,P_0)$ and $\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma_0,P_0)$ with respect to θ. For a generic (θ,γ,P), the criterion functions are

$$\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P) = \inf_{(F_x \in \mathcal{N}_{\delta_x})_{x\in\mathcal{X}}} \sum_x \mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{s.t. (41) holding at } (\theta,\gamma,P)\,,$$
$$\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P) = \sup_{(F_x \in \mathcal{N}_{\delta_x})_{x\in\mathcal{X}}} \sum_x \mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{s.t. (41) holding at } (\theta,\gamma,P)\,.$$

These profiled criterion functions have a dual formulation analogous to Proposition 2.1. Let $g(\cdot,x,\theta,\gamma)=(g_1(\cdot,x,\theta,\gamma),\ldots,g_4(\cdot,x,\theta,\gamma))$ denote the vector of moment functions evaluated at $X=x$. The remaining notation the same as in Proposition 2.1.

Proposition A.2 *Let Assumption Φ-conditional hold. If there exist distributions $F_x \in \mathcal{N}_{\delta_x}$ for all $x\in\mathcal{X}$ such that the moment conditions (41) hold at (θ,γ,P), then*

$$\begin{aligned}
&\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P)\\
&= \sup_{(\eta_x>0,\zeta_x\in\mathbb{R})_{x\in\mathcal{X}},\lambda\in\Lambda} \sum_x \left(-\eta_x \mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,x,\theta,\gamma)+\zeta_x+\lambda' g(U,x,\theta,\gamma)}{-\eta_x}\right)\right] - \eta_x\delta_x - \zeta_x - \lambda_{12}' P\right)\,, && (42)\\
&\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P)\\
&= \inf_{(\eta_x>0,\zeta_x\in\mathbb{R})_{x\in\mathcal{X}},\lambda\in\Lambda} \sum_x \left(\eta_x \mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,x,\theta,\gamma)-\zeta_x-\lambda' g(U,x,\theta,\gamma)}{\eta_x}\right)\right] + \eta_x\delta_x + \zeta_x + \lambda_{12}' P\right)\,. && (43)
\end{aligned}$$

Moreover, if the value of problem (42) is $+\infty$ (equivalently, if the value of problem (43) is $-\infty$), then for at least one $x\in\mathcal{X}$ there is no distribution in $\mathcal{N}_{\delta_x}$ under which the moment conditions (39) hold at (θ,γ,P).

Estimators $\hat{\underline{\kappa}}_{\boldsymbol{\delta}}$ and $\hat{\overline{\kappa}}_{\boldsymbol{\delta}}$ are again computed by minimizing and maximizing sample criterions with respect to θ, namely

$$\hat{\underline{K}}_{\boldsymbol{\delta}}(\theta) = \left[\begin{array}{l} \underline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P) \\ +\infty \end{array}\right. \quad , \quad \hat{\overline{K}}_{\boldsymbol{\delta}}(\theta) = \left[\begin{array}{ll} \overline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P) & \text{if } \Delta^\star_{sep}(\theta;\hat\gamma,\hat P)<0 \\ -\infty & \text{if } \Delta^\star_{sep}(\theta;\hat\gamma,\hat P)\geq 0, \end{array}\right.$$

where $\overline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P)$ and $\overline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P)$ denote the dual representations from Proposition A.2 evaluated at $(\hat\gamma,\hat P)$ with $\hat\gamma=(\hat{\tilde\gamma},\hat q)$ for estimators $\hat{\tilde\gamma}$ of $\tilde\gamma$ and $\hat q$ of q_0, and

$$\begin{aligned}
&\Delta^\star_{sep}(\theta;\gamma,P)\\
&= \sup_{\substack{(\eta_x\geq 0,\zeta_x\in\mathbb{R})_{x\in\mathcal{X}},\lambda\in\Lambda\\ \sum_{x\in\mathcal{X}}\eta_x\leq 1}} \left(-\sum_{x\in\mathcal{X}} \mathbb{E}^{F_*}\Big[(\eta_x\phi)^\star(-\zeta_x-\lambda_x' g(U,x,\theta,\gamma))\Big] - \eta_x\delta_x - \zeta_x\right) - \lambda_{12}' P\,.
\end{aligned}$$

Using similar arguments to Appendix F.3, the program $\Delta^\star_{sep}$ may be shown to be the dual of

$$\inf_{t\in\mathbb{R},(F_x)_{x\in\mathcal{X}}} t \quad \text{s.t.} \quad D_\phi(F_x\|F_*) \leq \delta_x + t \text{ for each } x \in \mathcal{X} \text{ and (41) holding at } (\theta,\gamma,P).$$

If there exist F_x with $D_\phi(F_x\|F_*) < \delta_x$ for each x such that (41) holds at (θ,γ,P), the value of this program is negative and hence $\Delta^\star_{sep}(\theta;\gamma,P)$ is negative. Consistency of the estimators and an asymptotic distribution theory may be derived by a suitable modification of the arguments in Section 6.

A.3 Exchangeability

Exchangeability can be an attractive shape restriction to impose on F in certain settings. For instance, in the context of discrete choice (Example 2.1), exchangeability ensures that choice probabilities depend on the alternatives' deterministic components of utility but not their labeling. Exchangeability can be easily imposed using our procedure whenever F_* is exchangeable, which is often the case in applications.

Let $U = (U_1,\ldots,U_J)$ and let Π_J denote the set of all permutations ϖ of $\{1,\ldots,J\}$. Given any $\Pi \subseteq \Pi_J$ that forms a group under composition, we say that F is Π-*exchangeable* if $(M_{\varpi(1)},\ldots,M_{\varpi(J)}) \sim F$ for all $\varpi \in \Pi$. Special cases include (full) exchangeability with $\Pi = \Pi_J$, cyclic exchangeability with $\Pi = \Pi^c_J := \{\varpi^j_c : j = 0,\ldots,J-1\}$ where $\varpi^j_c(i) = (i+j)(\text{mod}J)+1$ (i.e., invariance under rotations), and dihedral exchangeability when $J\geq 3$, with $\Pi = \Pi^c_J \cup \{\varpi^j_r : j = 0,\ldots,J-1\}$ where $\varpi^j_r(i) = (J-i+j)(\text{mod}J)+1$ (i.e., invariance under rotations and reflections).

Each notion of exchangeability ensures the marginal distribution of each element of U is identical, but they have different implications for the joint distributions of elements of U. For instance, the distribution of (U_i,U_j) for $i\neq j$ depends on $i-j$ and $|i-j|$ under cyclic and dihedral exchangeability, respectively, but is independent of (i,j) under (full) exchangeability.

Let $\mathcal{N}^{ex}_\delta = \{F \in \mathcal{N}_\delta : F \text{ is } \Pi\text{-exchangeable}\}$. Similar to (8) and (9), we are interested in

$$\underline{\kappa}^{ex}_\delta = \inf_{\theta\in\Theta,F\in\mathcal{N}^{ex}_\delta} \mathbb{E}^F[k(U,\theta,\gamma_0)] \quad \text{subject to (1)}, \tag{44}$$

$$\overline{\kappa}^{ex}_\delta = \sup_{\theta\in\Theta,F\in\mathcal{N}^{ex}_\delta} \mathbb{E}^F[k(U,\theta,\gamma_0)] \quad \text{subject to (1)}. \tag{45}$$

We again write these as optimization problems in terms of profiled criterion functions:

$$\underline{K}^{ex}_\delta(\theta;\gamma,P) = \inf_{F\in\mathcal{N}^{ex}_\delta} \mathbb{E}^F[k(U,\theta,\gamma)] \quad \text{subject to (1) holding at } (\theta,\gamma,P)\,, \tag{46}$$

$$\overline{K}^{ex}_\delta(\theta;\gamma,P) = \sup_{F\in\mathcal{N}^{ex}_\delta} \mathbb{E}^F[k(U,\theta,\gamma)] \quad \text{subject to (1) holding at } (\theta,\gamma,P)\,. \tag{47}$$

The values of the programs (46) and (47) can be computed using finite-dimensional convex programs when F_* is Π-exchangeable. Identify each ϖ with its corresponding permutation matrix $M_\varpi \in \{0,1\}^{J\times J}$. Define

$$k^{ex}(U,\theta,\gamma) = \frac{1}{|\Pi|}\sum_{\varpi\in\Pi} k(M_\varpi U,\theta,\gamma)\,, \quad g^{ex}_j(U,\theta,\gamma) = \frac{1}{|\Pi|}\sum_{\varpi\in\Pi} g_j(M_\varpi U,\theta,\gamma)\,, \quad j=1,2,3,4,$$

and let $g^{ex} = (g^{ex}_1, g^{ex}_2, g^{ex}_3, g^{ex}_4)$. We have the following counterpart to Proposition 2.1.

Proposition A.3 *Let Assumption Φ hold, let F_* be Π-exchangeable, and let Condition S hold at (θ,γ,P) for the moments g^{ex}. If there exists a distribution $F \in \mathcal{N}^{ex}_\delta$ such that the moment conditions (1) hold at (θ,γ,P), then*

$$\underline{K}^{ex}_\delta(\theta;\gamma,P) = \sup_{\eta>0,\zeta\in\mathbb{R},\lambda\in\Lambda} -\eta\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k^{ex}(U,\theta,\gamma)+\zeta+\lambda' g^{ex}(U,\theta,\gamma)}{-\eta}\right)\right] - \eta\delta - \zeta - \lambda_{12}' P\,, \tag{48}$$

$$\overline{K}^{ex}_\delta(\theta;\gamma,P) = \inf_{\eta>0,\zeta\in\mathbb{R},\lambda\in\Lambda} \eta\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k^{ex}(U,\theta,\gamma)-\zeta-\lambda' g^{ex}(U,\theta,\gamma)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda_{12}' P\,. \tag{49}$$

Moreover, if the value of problem (48) is $+\infty$ (equivalently, if the value of problem (49) is $-\infty$), then there is no distribution in $\mathcal{N}^{ex}_\delta$ under which (1) holds at (θ,γ,P).

Remark A.1 If F is Π-exchangeable and satisfies (1), then it must also satisfy (1) under all $|\Pi|$ permutations of the elements of U. The moment conditions imposed in the inner optimization are in fact

$$\begin{aligned} &\mathbb{E}^F[g_1(M_\varpi U,\theta,\gamma_0)] \le P_{10}, \quad &&\mathbb{E}^F[g_2(M_\varpi U,\theta,\gamma_0)] = P_{20},\\ &\mathbb{E}^F[g_3(M_\varpi U,\theta,\gamma_0)] \le 0, &&\mathbb{E}^F[g_4(M_\varpi U,\theta,\gamma_0)] = 0, \end{aligned} \tag{50}$$

for all permutations $\varpi\in\Pi$, which yields a total of $|\Pi|\times d$ moment conditions ($J!\times d$ for full exchangeability, and $J\times d$ and $2J\times d$ for cyclic and dihedral exchangeability, respectively, when $J\ge 3$). In principle one could include the full set of $|\Pi|\times d$ moments under permutations in the problems (48) and (49). By Π-exchangeability of F_* and convexity of the objective, the multipliers on the moments (50) will be identical across all permutations. Therefore, it suffices to consider only the d averaged moments in g^{ex} rather than the full set of $|\Pi|$ permutations of the d moments, reducing the dimension of the inner optimization by a factor of $|\Pi|$.

Remark A.2 When Monte Carlo methods are used to compute the expectations, the empirical distribution of the random draws from F_* may not be Π-exchangeabile even though the distribution F_* from which they are drawn is Π-exchangeable. In this case, Π-exchangeability of the empirical distribution of the draws can be imposed by taking a sample from F_* and then concatenating this sample across each of its $|\Pi|$ permutations.

B Additional Results on Nonparametrically Identified Sets

In this section we present further details to supplement Section 2.4 in the main text. Theorem 2.1 follows largely from the three lemmas presented in this section.

We first characterize the behavior of $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ as the neighborhood size δ becomes large. Let $\mathcal{N}_\infty = \{F : D_\phi(F \| F_*) < \infty\}$. Define

$$\mathcal{K}_\infty = \{\mathbb{E}^F[k(U, \theta, \gamma_0)] : (1) \text{ holds at } (\theta, \gamma_0, P_0),\ \theta \in \Theta,\ F \in \mathcal{N}_\infty\}\,.$$

The inclusion $\mathcal{K}_\infty \subseteq \mathcal{K}$ holds because $\mathcal{N}_\infty \subseteq \mathcal{F}_\theta$ for each $\theta \in \Theta$ by Assumption Φ. Our first result is that the smallest and largest values of $\mathcal{K}_\infty$ are approached by $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ as the neighborhood size δ gets large.

Lemma B.1 *Let Assumption Φ hold. Then $\underline{\kappa}_\delta \to \inf \mathcal{K}_\infty$ and $\overline{\kappa}_\delta \to \sup \mathcal{K}_\infty$ as $\delta \to \infty$.*

Next, we characterize the smallest and largest elements of $\mathcal{K}_\infty$ using profiled optimization problems and derive their dual forms. Define the profiled criterion functions

$$\underline{K}_\infty(\theta; \gamma_0, P_0) = \inf_{F \in \mathcal{N}_\infty} \mathbb{E}^F[k(U, \theta, \gamma_0)] \quad \text{subject to (1) holding at } (\theta, F)\,, \tag{51}$$

$$\overline{K}_\infty(\theta; \gamma_0, P_0) = \sup_{F \in \mathcal{N}_\infty} \mathbb{E}^F[k(U, \theta, \gamma_0)] \quad \text{subject to (1) holding at } (\theta, F)\,, \tag{52}$$

which are the infinite-δ versions of (11) and (12). By definition, we have

$$\inf \mathcal{K}_\infty = \inf_{\theta \in \Theta} \underline{K}_\infty(\theta; \gamma_0, P_0)\,, \qquad\qquad \sup \mathcal{K}_\infty = \sup_{\theta \in \Theta} \overline{K}_\infty(\theta; \gamma_0, P_0)\,.$$

Let F_*-$\operatorname{ess\,inf}$ and F_*-$\operatorname{ess\,sup}$ denote essential infimum and supremum defined relative to the measure F_*.

Lemma B.2 *Let Assumption Φ hold and let Condition S hold at (θ, γ, P). Then:*

$$\underline{K}_\infty(\theta; \gamma, P) = \sup_{\lambda \in \Lambda : F_*\text{-}\operatorname{ess\,inf}(k(\cdot,\theta,\gamma) + \lambda' g(\cdot,\theta,\gamma)) > -\infty} \left(F_*\text{-}\operatorname{ess\,inf}(k(\cdot, \theta, \gamma) + \lambda' g(\cdot, \theta, \gamma)) - \lambda_{12}' P\right)\,,$$

$$\overline{K}_\infty(\theta; \gamma, P) = \inf_{\lambda \in \Lambda : F_*\text{-}\operatorname{ess\,sup}(k(\cdot,\theta,\gamma) - \lambda' g(\cdot,\theta,\gamma)) < +\infty} \left(F_*\text{-}\operatorname{ess\,sup}(k(\cdot, \theta, \gamma) - \lambda' g(\cdot, \theta, \gamma)) + \lambda_{12}' P\right)\,.$$

The programs characterizing $\underline{K}_\infty(\theta; \gamma_0, P_0)$ and $\overline{K}_\infty(\theta; \gamma_0, P_0)$ at any fixed θ are non-smooth max-min and min-max optimization problems. These optimization problems may be difficult to solve, especially when u is multivariate and enters k or g nonlinearly. Note in particular that the inner optimization over u will typically be non-convex. Constraining F to $\mathcal{N}_\delta$ with $\delta < +\infty$ replaces the inner optimization over u with a nonlinear expectation, and results in the smooth convex programs (13) and (14).

Finally, we characterize the smallest and largest elements of $\mathcal{K}$ using profiled optimization problems. Define

$$\underline{K}_{np}(\theta;\gamma_0,P_0) = \inf_{F\in\mathcal{F}_\theta} \mathbb{E}^F[k(U,\theta,\gamma_0)] \quad \text{subject to (1) holding at } (\theta,F)\,, \tag{53}$$

$$\overline{K}_{np}(\theta;\gamma_0,P_0) = \sup_{F\in\mathcal{F}_\theta} \mathbb{E}^F[k(U,\theta,\gamma_0)] \quad \text{subject to (1) holding at } (\theta,F)\,. \tag{54}$$

We then have

$$\inf\mathcal{K} = \inf_{\theta\in\Theta} \underline{K}_{np}(\theta;\gamma_0,P_0)\,, \qquad \sup\mathcal{K} = \sup_{\theta\in\Theta} \overline{K}_{np}(\theta;\gamma_0,P_0)\,.$$

As with the other cases, the criterion functions $\underline{K}_{np}$ and $\overline{K}_{np}$ have a dual representation. Given the different topological nature of $\mathcal{F}_\theta$, we require a slightly different constraint qualification:

Definition B.1 Condition S_{np} *holds at* (θ,γ,P) *if* $\vec{P} \in \mathrm{ri}(\{\mathbb{E}^F[g(U,\theta,\gamma)] : F\in\mathcal{F}_\theta\} + \mathcal{C})$.

Condition S_{np} is weaker than Condition S from Section 2.4 provided F_* and μ are mutually absolutely continuous (see Lemma E.6). Under this condition, the criterion functions $\underline{K}_{np}$ and $\overline{K}_{np}$ admit a dual representation:

Lemma B.3 *Let Condition S_{np} hold at (θ,γ_0,P_0) and let $\mu\text{-}\mathrm{ess\,sup}\,|k(\cdot,\theta,\gamma_0)|<\infty$. Then:*

$$\underline{K}_{np}(\theta;\gamma_0,P_0) = \sup_{\lambda\in\Lambda:\mu\text{-}\mathrm{ess\,inf}(k(\cdot,\theta,\gamma_0)+\lambda' g(\cdot,\theta,\gamma_0))>-\infty} \left(\mu\text{-}\mathrm{ess\,inf}(k(\cdot,\theta,\gamma_0)+\lambda' g(\cdot,\theta,\gamma_0)) - \lambda_{12}' P_0\right)\,,$$

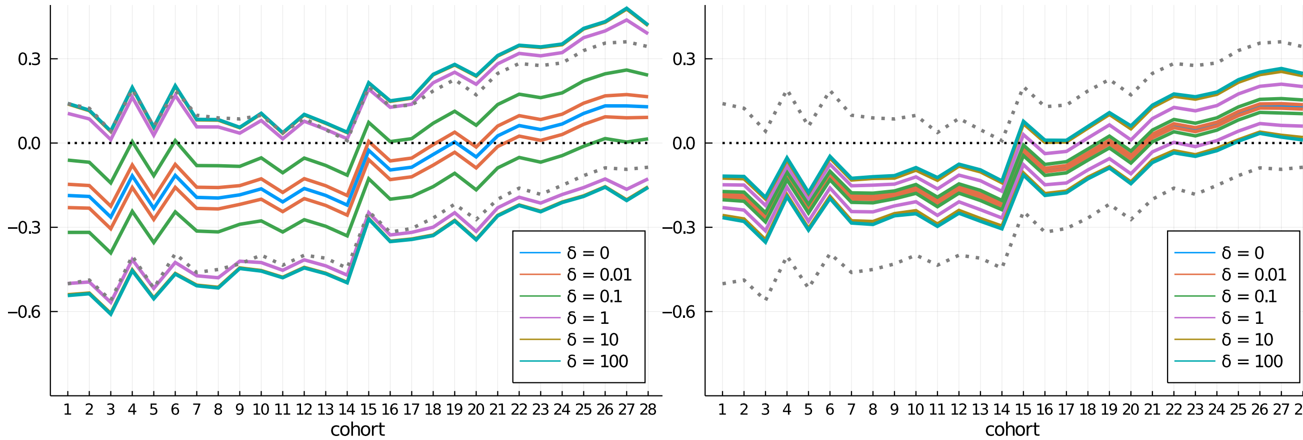
$$\overline{K}_{np}(\theta;\gamma_0,P_0) = \inf_{\lambda\in\Lambda:\mu\text{-}\mathrm{ess\,sup}(k(\cdot,\theta,\gamma_0)-\lambda' g(\cdot,\theta,\gamma_0))<+\infty} \left(\mu\text{-}\mathrm{ess\,sup}(k(\cdot,\theta,\gamma_0)-\lambda' g(\cdot,\theta,\gamma_0)) + \lambda_{12}' P_0\right)\,.$$

C Additional Details for Empirical Applications

This Appendix presents some additional details for the applications in Sections 5.1 and 5.2.

C.1 Marital College Premium

Additional results. Figure 6a plots lower and upper bounds on the SC to CG premium for our first implementation when structural parameters are held fixed at CSW's estimates $\hat{\theta}_*^c$ (computed under F_*) but F^c is allowed to vary. Figure 6b repeats this analysis for our second implementation imposing the exchangeability shape restriction. The fixed-θ nonparametric identified set for our second implementation is around half the width of the bounds $[\underline{\hat{\kappa}}_{0.01},\hat{\overline{\kappa}}_{0.01}]$ for our first implementation for late cohorts, and a fraction of the width for earlier cohorts.



(a) Implementation 1 (without exchangeability). (b) Implementation 2 (dihedral exchangeability).

Figure 6: Matching example: Bounds on the SC to CG premium for white women when F^c varies but structural parameters are held fixed at their estimates $\hat{\theta}^c_*$ under F_*. Grey dotted lines are the estimates $\underline{\hat{\kappa}}^c_{0.01}$ and $\hat{\bar{\kappa}}^c_{0.01}$ reported in Figure 1.

δ	0.01	0.10	1.00	10.0	100.0
Implementation 1	0.071	0.055	0.082	0.278	0.290
Implementation 2	0.197	0.271	0.363	0.313	0.588

Table 5: Matching example: Inner-optimization computation times (seconds) for $\underline{\hat{\kappa}}^1_\delta$. Implementations 1 and 2 use 50,000 and 180,000 Monte Carlo draws, respectively. All computations are performed in Julia version 1.5.3 and KNITRO 12.2.0 on a 2.7GHz MacBook Pro with 16GB memory.

Computation times. Table 5 reports average computation times for the inner optimization for evaluating the criterion function $\underline{\hat{K}}_\delta(\theta; \hat{P})$ for $\underline{\hat{\kappa}}^1_\delta$ at CSW's estimate $\hat{\theta}^1_*$ (the computation times are slightly different at different θ). Computation times increase somewhat with δ, as the Hessian becomes more ill-conditioned around the optimum for smaller δ. The computations reported in Table 5 initialize the solver at $\eta = 1$, $\zeta = 0$, and $\lambda = 0$. When embedded in the outer optimization with respect to θ, the inner-optimization computation times were reduced significantly by initializing at the (η, ζ, λ) solving the inner optimization at the previous value of θ. The outer optimization times varied with c, δ, and implementation but were typically solved in at most a few minutes (often 90 seconds or less).

Sensitivity to ϕ. Implementing the procedure with χ^2 and L^4 divergences produced near identical bounds for $\delta = 0.01$ and 0.1. With $\delta = 1$ and 10 the bounds with χ^2 divergence were slightly narrower than with hybrid divergence, but at most only around 10% narrower across cohorts for both premiums we analyze. The bounds with L^4 divergence were around 60% the width of the hybrid-neighborhood bounds for $\delta = 1$ and 10 across cohorts (L^4 divergence is stronger than χ^2 and hybrid divergence). The shapes of the sets with χ^2 and L^4 divergence were also similar to those reported for hybrid divergence. Overall, these results show that the conclusions we draw from our analysis are not sensitive to the choice of ϕ function.

δ	0.01	0.10	1.00	10.0	100.0
$\hat{\underline{\kappa}}_\delta$	0.119	0.163	0.203	0.184	0.730
$\hat{\overline{\kappa}}_\delta$	0.100	0.125	0.146	0.411	1.032

Table 6: Dynamic discrete choice example: Inner-optimization computation times (seconds) at the parameter values at which $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained. All computations are performed in Julia version 1.5.3 and KNITRO 12.2.0 on a 2.7GHz MacBook Pro with 16GB memory.

C.2 Welfare Analysis in a Rust Model

Computation times. Table 6 reports average computation times for the inner optimization for evaluating the criterion function $\hat{\underline{K}}_\delta(\theta; \hat{\gamma}, \hat{P})$ at the parameter values at which $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained. Here the computation times correspond to solving the minimum divergence problem $\Delta^\star(\theta; \hat{\gamma}, \hat{P})$ because k does not depend on u (cf. Section 2.5). Computation times are reported for the MPEC implementation, in which the inner optimization involves 75 moments (71 moments for CCPs and 4 location/scale normalizations). The computations reported in Table 6 initialize the solver at $\eta = 1$, $\zeta = 0$, and $\lambda = 0$. When embedded in the outer optimization with respect to θ, the inner-optimization computation times were reduced significantly by initializing at the (η, ζ, λ) solving the inner optimization at the previous value of θ. Here again the outer optimizations were typically solved in a few minutes or less.

Sensitivity to ϕ. Implementing the procedure with χ^2-divergence produced bounds that were at most 3% narrower and no wider than the bounds for hybrid divergence for all values of neighborhood size δ. Repeating the analysis with L^4-divergence, which is stronger than χ^2 and hybrid divergence, produced bounds that were 15-20% narrower than the bounds for hybrid divergence for values of δ up to $\delta = 1$ and at most 5% narrower than the hybrid divergence bounds for larger values of δ. As with the matching application, these results again show that the conclusions we draw from our analysis are not sensitive to the choice of ϕ function.

D Background Material on Orlicz Spaces

Our results rely on the theory of paired Orlicz spaces. We refer the reader to Krasnosel'skii and Rutickii (1961) for a textbook treatment. Here we summarize the relevant aspects that are used in our proofs. Define

$$\mathcal{L} = \{f : \mathcal{U} \to \mathbb{R} \text{ such that } \mathbb{E}^{F_*}[\phi(1 + c|f(U)|)] < \infty \text{ for some } c > 0\}$$
$$\mathcal{E} = \{f : \mathcal{U} \to \mathbb{R} \text{ such that } \mathbb{E}^{F_*}[\psi(c|f(U)|)] < \infty \text{ for all } c > 0\}\,.$$

The class $\mathcal{L}$ is an Orlicz class of functions corresponding to the function $x \mapsto \phi(1+|x|)$ whereas the class $\mathcal{E}$ is the Orlicz heart corresponding to $\psi(x) = \phi^\star(x) - x$ where $\phi^\star$ is the convex conjugate of ϕ. The condition $\lim_{x\to\infty} x\phi'(x)/\phi(x) < \infty$ in Assumption Φ(i) verifies the socalled Δ_2-condition in Krasnosel'skii and Rutickii (1961). The spaces $\mathcal{L}$ and $\mathcal{E}$ are separable Banach spaces when equipped with the norms

$$\|f\|_\phi = \inf_{c>0} \frac{1}{c}(1 + \mathbb{E}^{F_*}[\phi(1 + c|f(U)|)]) , \quad \text{and} \quad \|f\|_\psi = \inf_{c>0} \frac{1}{c}(1 + \mathbb{E}^{F_*}[\psi(c|f(U)|)]) ,$$

respectively (Krasnosel'skii and Rutickii, 1961, Chapter II, Section 10).[35] Given two functions ϕ_1, ϕ_2 satisfying Assumption Φ(i), write $\phi_1 \prec \phi_2$ if there exist positive constants c and x_0 such that $\phi_1(x) \leq \phi_2(cx)$ for all $x \geq x_0$. If $\phi_1 \prec \phi_2$ and $\phi_2 \prec \phi_1$ then ϕ_1 and ϕ_2 are said to be equivalent. Equivalent ϕ functions induce the same spaces $\mathcal{L}$ and $\mathcal{E}$ and their corresponding norms $\|\cdot\|_{\phi_1}$ and $\|\cdot\|_{\phi_2}$ are equivalent (Krasnosel'skii and Rutickii, 1961, Theorems 13.1 and 13.3). For example, the functions inducing hybrid and χ^2 divergence are equivalent.

A sequence $\{f_n\}_{n\geq 1} \subset \mathcal{L}$ is $\mathcal{E}$-weakly convergent if $\{\mathbb{E}^{F_*}[f_n(U)g(U)]\}_{n\geq 1}$ converges for each $g \in \mathcal{E}$. The space $\mathcal{L}$ is $\mathcal{E}$-weakly complete: any $\mathcal{E}$-weakly convergent sequence of functions $\{f_n\}_{n\geq 1} \subset \mathcal{L}$ has a unique limit, say $f^* \in \mathcal{L}$, for which

$$\lim_{n\to\infty} \mathbb{E}^{F_*}[f_n(U)g(U)] = \mathbb{E}^{F_*}[f^*(U)g(U)]$$

for each $g \in \mathcal{E}$; it is also $\mathcal{E}$-weakly compact: every $\|\cdot\|_\phi$-norm bounded sequence in $\mathcal{L}$ has an $\mathcal{E}$-weakly convergent subsequence (Krasnosel'skii and Rutickii, 1961, Theorem 14.4). A version of Hölder's inequality also holds:

$$|\mathbb{E}^{F_*}[f(U)g(U)]| \leq \|f\|_\phi \|g\|_\psi$$

for each $f \in \mathcal{L}$ and $g \in \mathcal{E}$ (Krasnosel'skii and Rutickii, 1961, Theorem 9.3).

The spaces $\mathcal{L}$ and $\mathcal{E}$ are paired locally convex topological vector spaces under the pairing

$$\langle f, g\rangle := \mathbb{E}^{F_*}[f(U)g(U)] , \quad f \in \mathcal{L},\ g \in \mathcal{E},$$

and suitable topologies on $\mathcal{L}$ and $\mathcal{E}$. Specifically, we equip $\mathcal{L}$ with the $\mathcal{E}$-weak topology, which ensures every continuous linear functional on $\mathcal{L}$ is representable in the form $\langle \cdot, g\rangle$ for $g \in \mathcal{E}$ (Krasnosel'skii and Rutickii, 1961, Theorem 14.7). Similarly, we equip $\mathcal{E}$ with the topology of $\mathcal{L}$-weak convergence, i.e., the sequence $\{g_n\}_{n\geq 1} \subset \mathcal{E}$ is $\mathcal{L}$-weakly convergent if $\{\mathbb{E}^{F_*}[f(U)g_n(U)]\}_{n\geq 1}$ converges for each $f \in \mathcal{L}$.[36] Under this topology, every continuous linear functional on $\mathcal{E}$ is representable in the form $\langle f, \cdot\rangle$ for $f \in \mathcal{L}$. Refer to Chapter 2, Section 14 of Krasnosel'skii and Rutickii (1961) and Section 3.3 of Komunjer and Ragusa (2016) for further details.

[35] In the notation of Krasnosel'skii and Rutickii (1961), our space $\mathcal{L}$ is the space $L_M^\star$ with $M(x) = \phi(1 + x)$ and the space $\mathcal{E}$ is the space E_N with $N(x) = \psi(x)$. As ϕ satisfies the Δ_2-condition, we have $L_M^\star = L_M = E_M$.

[36] In the notation of Krasnosel'skii and Rutickii (1961), this is equivalent to E_M-weak convergence (since $\mathcal{E}_M = L_M^\star$ by virtue of the Δ_2 condition) for functions in $E_N \subseteq L_N^\star$, with $M(x) = \phi(1 + x)$ and $N(x) = \psi(x)$.

We close this section by noting three useful results, the first of which is stated informally on p. 961 of Komunjer and Ragusa (2016).

Lemma D.1 *Under Assumption Φ(i), the functional $m \mapsto \mathbb{E}^{F_*}[\phi(m(U))]$ is l.s.c. on $\mathcal{L}$ in the $\mathcal{E}$-weak topology.*

Lemma D.2 *Under Assumption Φ(i), if $\mathbb{E}^{F_*}[\phi(m(U))] \leq \delta$ then $\|m\|_\phi \leq 2 + \phi(2) + \delta$.*

Let $\mathcal{L}_+ := \{m \in \mathcal{L} : m \geq 0 \ (F_*\text{-a.e.})\}$ denote the cone of non-negative functions in $\mathcal{L}$.

Lemma D.3 *Under Assumption Φ(i), $\mathbb{E}^{F_*}[\phi(m(U))] < \infty$ if and only if $m \in \mathcal{L}_+$.*

E Proof of Main Results

Throughout the proofs, we abbreviate upper-semicontinuous and upper-semicontinuity to u.s.c. and lower-semicontinuous and lower-semicontinuity to l.s.c.

E.1 Preliminary Results

We first present some preliminary results on the derivation of the dual formulation and verification of the constraint qualification conditions.

We derive the dual for $\underline{K}_\delta(\theta;\gamma,P)$; the derivation of the dual program for $\overline{K}_\delta$ follows similarly, replacing k with $-k$. Fix any $\theta \in \Theta$ and $\gamma \in \Gamma$. We drop dependence of $k(u,\theta,\gamma)$ and $g(u,\theta,\gamma)$ on (θ,γ) to simplify notation. For $\underline{K}_\delta$, the problem we wish to study is

$$\inf_F \mathbb{E}^F[k(U)] \quad \text{subject to} \quad D_\phi(F\|F_*) \leq \delta\,,\ \mathbb{E}^F[g_1(U)] \leq P_1\,,\ \ldots,\ \mathbb{E}^F[g_4(U)] = 0\,. \tag{55}$$

We apply duality theory as exposited in Chapter 2.5 of Bonnans and Shapiro (2000). We identify each distribution F for which $D_\phi(F\|F_*) < \infty$ with its Radon–Nikodym derivative $m = \frac{\mathrm{d}F}{\mathrm{d}F_*} \in \mathcal{L}$ (see Appendix D). We also pair the space $\mathcal{L}$ with the space $\mathcal{E}$, equipping both with the topologies described in Appendix D and using the pairing $\langle\cdot,\cdot\rangle$ defined therein.

Define the function $\varphi : \mathcal{L} \times \mathbb{R}^{d+2} \to \mathbb{R} \cup \{+\infty\}$ by

$$\varphi(m,y) = \langle m,k\rangle + \mathbb{I}_C\left(Q_\phi(m) - \delta + y_1, \langle m,1\rangle - 1 + y_2, \langle m,g\rangle - \vec{P} + y_3\right),$$

where $y = (y_1,y_2,y_3) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^d$, $\vec{P} = (P, 0_{d_3+d_4})$, $Q_\phi(m) = \mathbb{E}^{F_*}[\phi(m(U))]$,

$$\langle m,k\rangle = \mathbb{E}^{F_*}[m(U)k(U)]\,, \qquad \langle m,1\rangle = \mathbb{E}^{F_*}[m(U)]\,, \qquad \langle m,g\rangle = \mathbb{E}^{F_*}[m(U)g(U)]\,,$$

and $\mathbb{I}_C : \mathbb{R}^{d+2} \to \mathbb{R} \cup \{+\infty\}$ is given by

$$\mathbb{I}_C(y_1, y_2, y_3) = \left[\begin{array}{ll} 0 & \text{if } y_1 \le 0,\ y_2 = 0, \text{ and } y_3 \in \mathbb{R}^{d_1}_- \times \{0\}^{d_2} \times \mathbb{R}^{d_3}_- \times \{0\}^{d_4}\,, \\ +\infty & \text{otherwise}\,. \end{array} \right.$$

For any $y \in \mathbb{R}^{d+2}$, define the primal problem

$$\min_{m \in \mathcal{L}} \varphi(m, y) \tag{P$_y$}$$

and let $v(y) = \inf_{m \in \mathcal{L}} \varphi(m, y)$ denote its value. Then $v(0)$ is the value of problem (55).

We first establish some facts about φ and v. A convex function $f : X \to \mathbb{R} \cup \{+\infty\}$ is said to be *proper* if $f(x) > -\infty$ for all $x \in X$ and $f(x) < +\infty$ for some $x \in X$. The *effective domain* of f is $\operatorname{dom} f = \{x \in X : f(x) < +\infty\}$.

Lemma E.1 *Under Assumption Φ, the function φ is proper and convex.*

Proof of Lemma E.1. First note that $|\langle m, k\rangle| < +\infty$ for any $m \in \mathcal{L}$ by Hölder's inequality for Orlicz spaces (see Appendix D) and Assumption Φ(ii). It follows that $\varphi(m, y) > -\infty$ for all $m \in \mathcal{L}$ and $y \in \mathbb{R}^{d+2}$. Now take any $m \in \mathcal{L}_+$. Then $Q_\phi(m) < +\infty$ by Lemma D.3. Setting $y_1 = \delta - Q_\phi(m)$, $y_2 = 1 - \langle m, 1\rangle$ and $y_3 = \vec{P} - \langle m, g\rangle$ ensures $\mathbb{I}_C(Q_\phi(m) - \delta + y_1, \langle m, 1\rangle + y_2, \langle m, g\rangle - \vec{P} + y_3) = 0$, hence $\varphi(m, y) < +\infty$. This shows that φ is proper. Convexity of φ now follows from convexity of $m \mapsto Q_\phi(m)$ and convexity of $\operatorname{dom} \mathbb{I}_C$. ∎

Recall $\mathcal{C} = \mathbb{R}^{d_1}_+ \times \{0_{d_2}\} \times \mathbb{R}^{d_3}_+ \times \{0_{d_4}\}$. Let

$$\mathcal{Y} = \left\{ \begin{pmatrix} \delta - Q_\phi(m) \\ 1 - \langle m, 1\rangle \\ \vec{P} - \langle m, g\rangle \end{pmatrix} : m \in \mathcal{L}_+ \right\} - \mathbb{R}_+ \times \{0\} \times \mathcal{C}\,.$$

Lemma E.2 *Under Assumption Φ, (i) the function v is proper, convex, l.s.c., and* $\operatorname{dom} v = \mathcal{Y}$, *and (ii) a solution to the primal problem* (P$_y$) *exists for each $y \in \mathcal{Y}$.*

Proof of Lemma E.2. Convexity of v follows from convexity of φ established in Lemma E.1; see, e.g., Proposition 2.143 of Bonnans and Shapiro (2000).

The set $\mathcal{Y}$ is the set of all y for which there exists a $m_y \in \mathcal{L}$ for which

$$\mathbb{I}_C\left(Q_\phi(m_y) - \delta + y_1, \langle m_y, 1\rangle - 1 + y_2, \langle m_y, g\rangle - \vec{P} + y_3\right) < +\infty\,.$$

We have $|\langle m, k\rangle| < +\infty$ for any $m \in \mathcal{L}$ by Hölder's inequality (see Appendix D) and Assumption Φ(ii), and so $\varphi(m_y, y) < +\infty$ and hence $v(y) < +\infty$. Conversely, if $y \notin \mathcal{Y}$ then $\varphi(m, y) = +\infty$ for all $m \in \mathcal{L}$, and so $v(y) = +\infty$.

To see v is proper, take any $u \in \operatorname{dom} v$ and let y_1 denote its first element. Then

$$v(y) \geq \inf\{\langle k, m\rangle : m \in \mathcal{L}, Q_\phi(m) \leq \delta - y_1\},$$

and $\inf\{\langle k, m\rangle : m \in \mathcal{L}, Q_\phi(m) \leq \delta - y_1\} > -\infty$ by Hölder's inequality (see Appendix D), using Assumption Φ(ii) (which implies $\|k\|_\psi < +\infty$) and the fact that $\{m \in \mathcal{L} : Q_\phi(m) \leq \delta - y_1\}$ is $\|\cdot\|_\phi$-norm bounded (by Lemma D.2).

Before proving l.s.c. we first prove assertion (ii). Take any $y \in \mathcal{Y}$. Choose $\{m_n\}_{n\geq 1} \subset \mathcal{L}$ such that $\varphi(m_n, y) \downarrow v(y)$ as $n \to \infty$. As $Q_\phi(m_n) \leq \delta - y_1$ holds for each n, $\{m_n\}_{n\geq 1}$ is $\|\cdot\|_\phi$-norm bounded (by Lemma D.2) and therefore has a $\mathcal{E}$-weakly convergent subsequence $\{m_{n_l}\}_{l\geq 1}$ (see Appendix D). Let $m_0 \in \mathcal{L}$ denote the $\mathcal{E}$-weak limit. Under Assumption Φ, we have both $\lim_{l\to\infty}\langle m_{n_l}, 1\rangle = \langle m_0, 1\rangle$ and $\lim_{l\to\infty}\langle m_{n_l}, g\rangle = \langle m_0, g\rangle$ by the definition of $\mathcal{E}$-weak convergence, and also $\delta - y_1 \geq \liminf_{l\to\infty} Q(m_{n_l}) \geq Q(m_0)$ by Lemma D.1. It follows that

$$\mathbb{I}_C\left(Q_\phi(m_0) - \delta + y_1, \langle m_0, 1\rangle - 1 + y_2, \langle m_0, g\rangle - \vec{P} + y_3\right) = 0,$$

so m_0 is feasible for the primal problem. Moreover, by $\mathcal{E}$-weak convergence we also have that $v(y) = \lim_{l\to\infty}\langle m_{n_l}, k\rangle = \langle m_0, k\rangle$. Therefore, m_0 solves the primal problem (P_y).

To prove l.s.c., take any $y \in \mathcal{Y}$. Take $\{y_n\}_{n\geq 1} \subset \mathcal{Y}$ converging to y. By the argument used to establish properness, we have that $\{v(y_n)\}_{n\geq 1}$ is bounded and therefore has a subsequence $\{y_{n_l}\}_{l\geq 1}$ converging to $\liminf_{n\to\infty} v(y_n)$. Let m_{n_l} solve the primal problem for each y_{n_l}. The sequence $\{m_{n_l}\}_{l\geq 1}$ is $\|\cdot\|_\phi$-norm bounded (by Lemma D.2) and hence, taking a further subsequence if necessary, has an $\mathcal{E}$-weak limit $m_0 \in \mathcal{L}$ (see Appendix D). By similar arguments to the above, we may deduce that m_0 is feasible for the primal problem at y. It then follows again by $\mathcal{E}$-weak convergence that

$$\liminf_{n\to\infty} v(y_n) = \lim_{l\to\infty} v(y_{n_l}) = \lim_{l\to\infty}\langle m_{n_l}, k\rangle = \langle m_0, k\rangle \geq v(y),$$

as required. ■

The dual problem of (P_y) is (Bonnans and Shapiro, 2000, p. 96)

$$\max_{y^\star \in \mathbb{R}^{d+2}} y' y^\star - \varphi^\star(0, y^\star), \tag{D_y}$$

where $\varphi^\star : \mathcal{E} \times \mathbb{R}^{d+2} \to \mathbb{R} \cup \{+\infty\}$ is the convex conjugate of φ:

$$\varphi^\star(m^\star, y^\star) = \sup_{(m,y)\in\mathcal{L}\times\mathbb{R}^{d+2}} \left(\langle m, m^\star\rangle + y' y^\star - \varphi(m, y)\right),$$

and $y^\star = (y_1^\star, y_2^\star, y_3^\star) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$. By direct calculation,

$$\begin{aligned}\varphi^\star(0, y^\star) &= \sup_{(m,y)\in\mathcal{L}\times\mathbb{R}^{d+2}} \Big(y'y^\star - \langle m, k\rangle - \mathbb{I}_C\Big(Q_\phi(m) - \delta + y_1, \langle m, 1\rangle - 1 + y_2, \langle m, g\rangle - \vec{P} + y_3\Big)\Big) \\ &= \sup_{m\in\mathcal{L}} \Bigg(- y_1^\star(Q_\phi(m) - \delta) - y_2^\star(\langle m, 1\rangle - 1) - y_3^{\star\prime}\Big(\langle m, g\rangle - \vec{P}\Big) - \langle m, k\rangle\Bigg) + \mathbb{I}_{C^o}(y^\star)\end{aligned}$$

where $C^o = \mathbb{R}_+ \times \mathbb{R} \times \Lambda$ is the polar cone of $C := \operatorname{dom} \mathbb{I}_C$.

Write any $y^\star \in C^o$ as $y^\star = (\eta, \zeta, \lambda) \in \mathbb{R}_+ \times \mathbb{R} \times \Lambda$. We then have

$$\begin{aligned}\varphi^\star(0, (\eta, \zeta, \lambda)) &= \sup_{m\in\mathcal{L}} \Bigg(- \eta Q_\phi(m) - \zeta\langle m, 1\rangle - \lambda'\langle m, g\rangle - \langle m, k\rangle\Bigg) + \eta\delta + \zeta + \lambda_{12}'P \\ &= \sup_{m\in\mathcal{L}} \mathbb{E}^{F_*}\big[m(U)(-k(U) - \zeta - \lambda'g(U)) - \eta\phi(m(U))\big] + \eta\delta + \zeta + \lambda_{12}'P\,. \quad (56)\end{aligned}$$

As $\mathcal{L}$ is decomposable (Rockafellar and Wets, 1998, Definition 14.59 and Theorem 14.60), we may bring the supremum inside the expectation and optimize pointwise to obtain

$$\varphi^\star(0, (\eta, \zeta, \lambda)) = \mathbb{E}^{F_*}\Big[(\eta\phi)^\star(-k(U) - \zeta - \lambda'g(U))\Big] + \eta\delta + \zeta + \lambda_{12}'P$$

provided $(\eta, \zeta, \lambda) \in C^o$, where

$$(\eta\phi)^\star(x) = \sup_{t\geq 0:\eta\phi(t)<+\infty} (tx - \eta\phi(t)) = \begin{bmatrix} \eta\phi^\star(x/\eta) & \text{if } \eta > 0 \\ 0 & \text{if } \eta = 0 \text{ and } x \leq 0 \\ +\infty & \text{if } \eta = 0 \text{ and } x > 0 \end{bmatrix}$$

is the convex conjugate of $x \mapsto \eta \times \phi(x)$. As $\langle y, y^\star\rangle - \varphi^\star(0, y^\star) = -\infty$ whenever $y^\star \notin C^o$, the dual problem (D_y) can therefore be expressed as

$$\max_{\eta\geq 0,\zeta\in\mathbb{R},\lambda\in\Lambda} \eta y_1 + \zeta y_2 + \lambda' y_3 - \mathbb{E}^{F_*}\Big[(\eta\phi)^\star(-k(U) - \zeta - \lambda'g(U))\Big] - \eta\delta - \zeta - \lambda_{12}'P\,.$$

In particular, the dual of (55), which corresponds to the dual (D_y) of (P_y) with $y = 0$, is

$$\max_{\eta\geq 0,\zeta\in\mathbb{R},\lambda\in\Lambda} -\mathbb{E}^{F_*}\Big[(\eta\phi)^\star(-k(U) - \zeta - \lambda'g(U))\Big] - \eta\delta - \zeta - \lambda_{12}'P\,. \quad (57)$$

Note the value of the dual problem (D_y) is the biconjugate $v^{\star\star}(y)$. The following result establishes conditions for equality of the primal and dual problems in displays (55) and (57).

Lemma E.3 *Let Assumption Φ hold. Then: (i) if $0 \in \mathcal{Y}$ then the value of the primal and dual problems (55) and (57) are equal; (ii) if $0 \in \operatorname{ri}(\mathcal{Y})$ then the set of solutions of the dual problem is nonempty and convex; and (iii) if $0 \in \operatorname{int}(\mathcal{Y})$ then the set of dual solutions is also compact.*

Proof of Lemma E.3. Part (i) follows by Lemma E.2(i) and the discussion following Theorem 2.144 on p. 98 of Bonnans and Shapiro (2000).

For part (ii), non-emptiness of the set of dual solutions follows by Propositions 2.147 and 2.148(iii) of Bonnans and Shapiro (2000), noting that v is convex by Lemma E.2 and $v(0)$ is finite because v is proper by Lemma E.2 and $0 \in \operatorname{dom} v \equiv \mathcal{Y}$ by Assumption. Convexity of the set of dual solutions follows by noting that, in view of (D_y) and (56), the dual objective is the pointwise infimum of affine functions of (η, ζ, λ), and is therefore concave and u.s.c.

Part (iii) follows from Theorem 2.151 and Proposition 2.152 of Bonnans and Shapiro (2000). ■

Recall Condition S and the set $\mathcal{C}$ from Section 2.4 and Condition S' from Section 6. Define

$$\begin{aligned} \mathcal{Y}_1 &= \left\{ \vec{P} - \langle m, g \rangle : m \in \mathcal{L}_+, \langle m, 1 \rangle = 1 \right\} - \mathcal{C}\,, \\ \mathcal{Y}_2 &= \left\{ \left(\begin{array}{c} 1 - \langle m, 1 \rangle \\ \vec{P} - \langle m, g \rangle \end{array} \right) : m \in \mathcal{L}_+ \right\} - \{0\} \times \mathcal{C}\,. \end{aligned} \tag{58}$$

Let 0 denote a vector of zeros whose dimension is determined by the context.

Lemma E.4 *Let Assumption Φ hold and Condition S hold at (θ, γ, P). Then: (i) $0 \in \operatorname{ri}(\mathcal{Y}_1)$; (ii) $0 \in \operatorname{ri}(\mathcal{Y}_2)$; and (iii) if there exists F with $D_\phi(F \| F_*) < \delta$ s.t. the conditions in (55) hold at θ, then $0 \in \operatorname{ri}(\mathcal{Y})$. Moreover, if Condition S' holds at (θ, γ, P) then "relative interior" can be replaced with "interior" in parts (i)–(iii).*

Proof of Lemma E.4. In view of Lemma D.3, identify each $F \in \mathcal{N}_\infty := \{F : D_\phi(F \| F_*) < \infty\}$ with its Radon–Nikodym derivative with respect to F_*, say $m \in \mathcal{L}$. Part (i) follows by noting

$$\vec{P} \in \operatorname{ri}(\{\mathbb{E}^F[g(U)] : F \in \mathcal{N}_\infty\} + \mathcal{C}) \iff \vec{P} \in \operatorname{ri}(\{\langle m, g \rangle : m \in \mathcal{L}_+, \langle m, 1 \rangle = 1\} + \mathcal{C}) \iff 0 \in \operatorname{ri}(\mathcal{Y}_1)\,.$$

To prove part (ii), note that showing $0 \in \operatorname{ri}(\mathcal{Y}_2)$ is equivalent to showing $(1, \vec{P}) \in \operatorname{ri}(\mathcal{V}_2)$, where

$$\mathcal{V}_2 = \left\{ \left(\begin{array}{c} \langle m, 1 \rangle \\ \langle m, g \rangle \end{array} \right) : m \in \mathcal{L}_+ \right\} + \{0\} \times \mathcal{C} = \operatorname{cone}(\mathcal{V}_1) + \{0\} \times \mathcal{C} = \operatorname{cone}\left(\mathcal{V}_1 + \{0\} \times \mathcal{C}\right)\,,$$

$\mathcal{V}_1 = \{(1, \langle m, g \rangle) : m \in \mathcal{L}_+, \langle m, 1 \rangle = 1\}$, and $\operatorname{cone}(A) = \{ta : a \in A, t \geq 0\}$. But we have $\operatorname{ri}(\mathcal{V}_2) = \{tv : v \in \operatorname{ri}(\mathcal{V}_1 + \{0\} \times \mathcal{C}), t > 0\}$. Moreover, $(1, \vec{P}) \in \operatorname{ri}(\mathcal{V}_1 + \{0\} \times \mathcal{C})$ by Condition S because $\operatorname{ri}(\{1\} \times A) = \{1\} \times \operatorname{ri}(A)$. Therefore, $(1, \vec{P}) \in \operatorname{ri}(\mathcal{V}_2)$.

To prove part (iii), note that showing $0 \in \operatorname{ri}(\mathcal{Y})$ is equivalent to showing $(\delta, 1, \vec{P}) \in \operatorname{ri}(\mathcal{V}_3)$, where

$$\mathcal{V}_3 = \left\{ \left(\begin{array}{c} Q_\phi(m) \\ \langle m, 1 \rangle \\ \langle m, g \rangle \end{array} \right) : m \in \mathcal{L}_+ \right\} + \mathbb{R}_+ \times \{0\} \times \mathcal{C}\,.$$

It suffices to show that for every $v \in \mathcal{V}_3$ there exists some $t > 1$ such that $t(\delta, 1, \vec{P}) + (1-t)v \in \mathcal{V}_3$ (Rockafellar, 1970, Theorem 6.4). Take any $v \in \mathcal{V}_3$. Then we may write $v = (v_1, v_2) \in \mathbb{R}_+ \times \mathcal{V}_2$. By part (ii) and Theorem 6.4 of Rockafellar (1970) that there exists $s > 1$, $m_v \in \mathcal{L}_+$, and $c_3 \in \mathcal{C}$ such that

$$s \begin{pmatrix} 1 \\ \vec{P} \end{pmatrix} + (1-s)v_2 = \begin{pmatrix} \langle m_v, 1 \rangle \\ \langle m_v, g \rangle \end{pmatrix} + \begin{pmatrix} 0 \\ c_3 \end{pmatrix} . \tag{59}$$

By assumption, there exists $F \in \mathcal{N}_\delta$ with $D_\phi(F \| F_*) < \delta$ such that the moment conditions in (55) hold at θ. Let $\tilde{m}$ denote the Radon–Nikodym derivative of such an F. Then for any $\tau \in (0,1)$, setting $m_\tau = \tau m_v + (1-\tau)\tilde{m}$, we have $\langle m_\tau, 1 \rangle = \tau \langle m_v, 1 \rangle + (1-\tau)$ and $\langle m_\tau, g \rangle = \tau \langle m_v, g \rangle + (1-\tau)(\vec{P} - \tilde{c})$ for some $\tilde{c} \in \mathcal{C}$. But then

$$\begin{pmatrix} \langle m_v, 1 \rangle \\ \langle m_v, g \rangle \end{pmatrix} = \frac{1}{\tau} \begin{pmatrix} \langle m_\tau, 1 \rangle \\ \langle m_\tau, g \rangle \end{pmatrix} - \frac{1-\tau}{\tau} \begin{pmatrix} 1 \\ \vec{P} - \tilde{c} \end{pmatrix} . \tag{60}$$

Substituting (60) into (59) yields

$$(1 + \tau(s-1)) \begin{pmatrix} 1 \\ \vec{P} \end{pmatrix} - \tau(s-1)v_2 = \begin{pmatrix} \langle m_\tau, 1 \rangle \\ \langle m_\tau, g \rangle \end{pmatrix} + \begin{pmatrix} 0 \\ \tau \bar{c}_3 + (1-\tau)\tilde{c} \end{pmatrix} .$$

Note that $Q_\phi(m_\tau)$ can be made arbitrarily close to $Q_\phi(\tilde{m}) < \delta$ by choosing τ arbitrarily small. Setting $t = 1 + \tau(s-1)$ with τ sufficiently small that $t\delta + (1-t)v_1 \geq Q_\phi(m_\tau)$, we may write

$$t \begin{pmatrix} \delta \\ 1 \\ \vec{P} \end{pmatrix} + (1-t)v = \begin{pmatrix} Q_\phi(m_\tau) \\ \langle m_\tau, 1 \rangle \\ \langle m_\tau, g \rangle \end{pmatrix} + \begin{pmatrix} c_1 \\ 0 \\ \tau c_3 + (1-\tau)\tilde{c} \end{pmatrix}$$

for some $c_1 \geq 0$. As the right-hand side belongs to $\mathcal{V}_3$, this completes the proof of part (iii).

Now suppose Condition S' holds. Part (i) holds with "interior" by definition of Condition S'. For part (ii) with "interior", it suffices to show that $\mathcal{Y}_2$ has positive volume, in which case its relative interior and interior coincide and the result follows by part (ii) above. A sufficient condition is that the functions in g and a function that is constant F_*-a.e. are not collinear F_*-a.e. We prove this by contradiction. Suppose Condition S' holds but that there exists $0 \neq \lambda \in \mathbb{R}^d$ and $\zeta \in \mathbb{R}$ such that $\lambda'(g(u) - \vec{P}) = \zeta$ F_*-a.e. Then by Condition S', we have $\{\mathbb{E}^F[g(U)] - \vec{P} : D_\phi(F \| F_*) < \infty\}$ contains a ε-ball with center c_0 for some $c_0 \in \mathcal{C}$ and $\varepsilon > 0$. But then for any unit vector u we have $\zeta = \lambda' c_0 + \varepsilon \lambda' u$, a contradiction. Thus, part (ii) must hold with "interior" when Condition S' holds. For part (iii), note that $\mathcal{Y} \supseteq (\{\delta\} + \mathbb{R}_-) \times \mathcal{Y}_2$. Therefore, $\mathcal{Y}$ has positive volume as $\mathcal{Y}_2$ as positive volume, so its relative interior and interior coincide and part (iii) with "interior" follows similarly. ■

Lemma E.5 *Let Assumption Φ hold. Then: the dual programs of $\underline{K}_\delta(\theta; \gamma, P)$ and $\overline{K}_\delta(\theta; \gamma, P)$*

are

$$\underline{K}^\star_\delta(\theta;\gamma,P) = \sup_{\eta\geq 0,\zeta\in\mathbb{R},\lambda\in\Lambda} -\mathbb{E}^{F_*}\left[(\eta\phi)^\star(-k(U,\theta,\gamma)-\zeta-\lambda' g(U,\theta,\gamma))\right] - \eta\delta - \zeta - \lambda_{12}' P\,, \text{ and}$$
$$\overline{K}^\star_\delta(\theta;\gamma,P) = \inf_{\eta\geq 0,\zeta\in\mathbb{R},\lambda\in\Lambda} \mathbb{E}^{F_*}\left[(\eta\phi)^\star(k(U,\theta,\gamma)-\zeta-\lambda' g(U,\theta,\gamma))\right] + \eta\delta + \zeta + \lambda_{12}' P\,.$$

If there exists $F\in\mathcal{N}_\delta$ such that the moment conditions (1) hold at (θ,γ,P), then:

$$\underline{K}_\delta(\theta;\gamma,P) = \underline{K}^\star_\delta(\theta;\gamma,P)\,, \qquad \overline{K}_\delta(\theta;\gamma,P) = \overline{K}^\star_\delta(\theta;\gamma,P)\,,$$

and the supremum and infimum can be taken over $(\eta,\zeta,\lambda)\in(0,\infty)\times\mathbb{R}\times\Lambda$ in the definition of $\underline{K}^\star_\delta(\theta;\gamma,P)$ and $\overline{K}^\star_\delta(\theta;\gamma,P)$. Moreover, if Condition S holds and there exists $F\in\mathcal{N}_\delta$ with $D_\phi(F\|F_*)<\delta$ such that the moment conditions (1) hold under F at (θ,γ,P), then solutions to both dual problems (with $(\eta,\zeta,\lambda)\in\mathbb{R}_+\times\mathbb{R}\times\Lambda$) exist. Moreover, if Condition S' holds, then the set of dual solutions is compact.

Proof of Lemma E.5. We prove only the result for $\underline{K}_\delta$; the result for $\overline{K}_\delta$ follows similarly.

The dual program $\underline{K}^\star_\delta$ is derived in (57) above, which requires only Assumption Φ to hold. Equality of the primal and dual programs follows by Lemma E.3(i), noting that existence of $F\in\mathcal{N}_\delta$ such that (1) holds at (θ,γ,P) under F ensures that $0\in\mathcal{Y}$.

It remains to show that the supremum and infimum can be taken over $(\eta,\zeta,\lambda)\in(0,\infty)\times\mathbb{R}\times\Lambda$. In view of ($\mathrm{D}_y$) and (56), the dual objective function, say $\ell(\eta,\zeta,\lambda)$, is the pointwise infimum of affine functions of (η,ζ,λ), and is therefore concave and u.s.c. Note $\ell(\eta,\zeta,\lambda)<\infty$ for all $(\eta,\zeta,\lambda)\in\mathbb{R}_+\times\mathbb{R}\times\Lambda$ as the primal value and hence dual value is finite. If $\ell(0,\zeta,\lambda)=-\infty$ for all $\zeta\in\mathbb{R}$ and $\lambda\in\Lambda$, then restricting (η,ζ,λ) to $(0,\infty)\times\mathbb{R}\times\Lambda$ will not affect the dual value. Now suppose that there is some $(0,\zeta^*,\lambda^*)$ with $(\zeta^*,\lambda^*)\in\mathbb{R}\times\Lambda$ for which $\ell(\eta,\zeta^*,\lambda^*)>-\infty$. Then by u.s.c. and concavity, $\ell(\cdot,\zeta^*,\lambda^*)$ is continuous on $[0,\overline{\eta}]$ for $\overline{\eta}>0$ (Rockafellar, 1970, Theorem 10.2) hence $\lim_{\eta\downarrow 0}\ell(\eta,\zeta^*,\lambda^*)=\ell(0,\zeta^*,\lambda^*)$.

If Condition S holds and there exists $F\in\mathcal{N}_\delta$ with $D_\phi(F\|F_*)$ that satisfies the moment conditions, then $0\in\mathrm{ri}(\mathcal{Y})$ by Lemma E.4. Existence of a dual solution then follows by Lemmas E.3(ii). Compactness of the set of dual solutions under Condition S' follows similarly by Lemmas E.3(iii) and Lemma E.4. ■

E.2 Proofs for Section 2

Proof of Proposition 2.1. Follows by Lemma E.5 and $(\eta\phi)^\star(x)=\eta\phi^\star(x/\eta)$ for $\eta>0$. ■

The proof of Theorem 2.1 uses results from Appendix B derived under different constraint qualification conditions: Condition S from Section 2.4 and Condition S_{np} from Appendix B. We first present a lemma relating these conditions.

Lemma E.6 *Under Assumption Φ, if μ and F_* are mutually absolutely continuous and Condition S holds at (θ, γ, P), then Condition S_{np} also holds at (θ, γ, P).*

Proof of Lemma E.6. Assumption Φ implies $\mathcal{N}_\infty = \{F : D_\phi(F \| F_*) < \infty\} \subseteq \mathcal{F}_\theta$. Therefore,

$$\mathcal{G}(\theta, \gamma) := \{\mathbb{E}^F[g(U, \theta, \gamma)] : F \in \mathcal{N}_\infty\} \subseteq \{\mathbb{E}^F[g(U, \theta, \gamma)] : F \in \mathcal{F}_\theta\} =: \mathcal{G}_\theta(\gamma) \,.$$

Write $\mathcal{G}(\theta, \gamma) = \mathcal{G}_\infty$ and $\mathcal{G}_\theta(\gamma) = \mathcal{G}_\theta$ to simplify notation. By Corollary 6.6.2 of Rockafellar (1970), it suffices to show $\mathrm{ri}(\mathcal{G}_\infty) \subseteq \mathrm{ri}(\mathcal{G}_\theta)$. To this end, we shall actually show $\mathrm{ri}(\mathcal{G}_\infty) = \mathrm{ri}(\mathcal{G}_\theta)$. As $\mathrm{ri}(\mathcal{G}_\infty) \subseteq \mathcal{G}_\theta$, it suffices to show $\mathcal{G}_\theta \subseteq \mathrm{cl}(\mathcal{G}_\infty)$ (Hiriart-Urruty and Lemaréchal, 2001, Remark 2.1.9). For any $x \in \mathcal{G}_\theta$, we have $x = \mathbb{E}^F[g(U, \theta, \gamma)]$ for some $F \in \mathcal{F}_\theta$. As $F \ll \mu$ and F_* and μ are mutually absolutely continuous, F has a density, say m, with respect to F_*. For each $n \geq 1$, define

$$m_n(u) = \frac{m(u) \wedge n}{\int (m \wedge n) \, \mathrm{d}F_*} \,.$$

Each m_n is bounded and hence each measure F_n defined by $\mathrm{d}F_n = m_n \mathrm{d}F_*$ belongs to $\mathcal{N}_\infty$. By monotone convergence, we have $\mathcal{G}_\infty \ni \mathbb{E}^{F_n}[g(U, \theta, \gamma)] \to x$. Therefore, $x \in \mathrm{cl}(\mathcal{G}_\infty)$. ■

Proof of Theorem 2.1. We prove only the result for $\inf \mathcal{K}$; the result for $\sup \mathcal{K}$ follows similarly. First note

$$\inf \mathcal{K} = \inf_{\theta \in \Theta} \underline{K}_{np}(\theta; \gamma_0, P_0) = \inf_{\theta \in \Theta_I} \underline{K}_{np}(\theta; \gamma_0, P_0) \,,$$

where $\underline{K}_{np}$ is defined in (53). The first equality is by definition and the second equality holds because if $\theta \notin \Theta_I$, then there does not exist a distribution $F \in \mathcal{F}_\theta$ under which the moment conditions hold at (θ, γ_0, P_0) and consequently $\underline{K}_{np}(\theta; \gamma_0, P_0) = +\infty$. If $\theta \notin \Theta_I$, then there does not exist $F \in \mathcal{N}_\infty$ under which the moment conditions hold at (θ, γ_0, P_0) either because $\mathcal{N}_\infty \subseteq \mathcal{F}_\theta$. Therefore, $\underline{K}_\infty(\theta; \gamma_0, P_0) = +\infty$ in that case too. We therefore have

$$\inf \mathcal{K}_\infty = \inf_{\theta \in \Theta_I} \underline{K}_\infty(\theta; \gamma_0, P_0) \,.$$

As Condition S holds at (θ, γ_0, P_0) for each $\theta \in \Theta_I$ and $\mu \ll F_* \ll \mu$, Lemma E.6 implies that Condition S_{np} must also hold at (θ, γ_0, P_0) for each $\theta \in \Theta_I$. As μ and F_* are mutually absolutely continuous, the μ- and F_*-essential supremums of any function are equal. Therefore by Lemmas B.2 and B.3, we have $\underline{K}_\infty(\theta; \gamma_0, P_0) = \underline{K}_{np}(\theta; \gamma_0, P_0)$ for each $\theta \in \Theta_I$. The result now follows by Lemma B.1. ■

E.3 Proofs for Section 3

Proof of Proposition 3.1. We prove the result for $\underline{K}_\delta$; the proof for $\overline{K}_\delta$ follows similarly.

Consider two programs:

$$v^A := \inf_{\theta\in\Theta, F\in\mathcal{N}_\delta} \mathbb{E}^F[k(U,\theta,\gamma)] \quad \text{subject to (1) holding at } (\theta,\gamma,P)\,, \tag{Program A}$$

$$v^B := \inf_{\theta\in\Theta} \mathbb{E}^{\underline{F}_{\delta,\theta}}[k(U,\theta,\gamma)] \quad \text{subject to } \mathbb{E}^{\underline{F}_{\delta,\theta}}[g_{4e}(U,\theta,\gamma)] = 0\,, \tag{Program B}$$

where $\underline{F}_{\delta,\theta}$ solves

$$\inf_{F\in\mathcal{N}_\delta} \mathbb{E}^F[k(U,\theta,\gamma)] \quad \text{subject to (18) holding at } (\theta,\gamma,P)\,,$$

and $v^B = +\infty$ if there is no solution to this problem. Program A is the approach described in Section 2 whereas Program B is equivalent to our MPEC implementation.

The inequality $v^A \le v^B$ is trivial if $v^B = +\infty$. If v^B is finite, for any $\varepsilon > 0$ there exists $\theta^B_\varepsilon \in \Theta$ for which $\mathbb{E}^{\underline{F}_{\delta,\theta^B_\varepsilon}}[k(U,\theta^B_\varepsilon,\gamma)] \le v^B + \varepsilon$ and $\mathbb{E}^{\underline{F}_{\delta,\theta^B_\varepsilon}}[g_{4e}(U,\theta^B_\varepsilon,\gamma)] = 0$ where $\underline{F}_{\delta,\theta^B_\varepsilon}$ is well defined by Lemma E.2(ii). As $(\theta^B_\varepsilon, \underline{F}_{\delta,\theta^B_\varepsilon})$ are feasible for Program A, we have $v^A \le v^B + \varepsilon$. As ε is arbitrary, we have $v^A \le v^B$.

A similar argument applies when $v^B = -\infty$: for any $n \in \mathbb{N}$ there exists $\theta^B_n \in \Theta$ for which $\mathbb{E}^{\underline{F}_{\delta,\theta^B_n}}[k(U,\theta^B_\varepsilon,\gamma)] \le -n$ and $\mathbb{E}^{\underline{F}_{\delta,\theta^B_n}}[g_{4e}(U,\theta^B_\varepsilon,\gamma)] = 0$, where the distribution $\underline{F}_{\delta,\theta^B_n}$ is well defined by Lemma E.2(ii). As $(\theta^B_n, \underline{F}_{\delta,\theta^B_n})$ are feasible for Program A, we have $v^A \le -n$. As this is true for all $n \in \mathbb{N}$, we have $v^A = v^B$.

The inequality $v^B \le v^A$ holds trivially if $v^A = +\infty$. If v^A is finite, rewrite Program B as

$$\inf_{\kappa\in\mathbb{R},\theta\in\Theta} \kappa \quad \text{subject to } \mathbb{E}^{\underline{F}_{\delta,\theta,\kappa}}[g_{4e}(U,\theta,\gamma)] = 0\,,$$

where $\underline{F}_{\delta,\theta,\kappa}$ solves the feasibility program

$$\inf_{F\in\mathcal{N}_\delta} 0 \quad \text{subject to (18) and } \mathbb{E}^F[k(U,\theta,\gamma)] = \kappa \text{ holding at } (\theta,\gamma,P). \tag{61}$$

For any $\varepsilon > 0$ there exists $\theta^A_\varepsilon \in \Theta$ and $F^A_\varepsilon \in \mathcal{N}_\delta$ such that the constraints in Program A are satisfied, i.e. $\mathbb{E}^{F^A_\varepsilon}[g_1(U,\theta^A_\varepsilon,\gamma)] \le P_1, \ldots, \mathbb{E}^{F^A_\varepsilon}[g_4(U,\theta^A_\varepsilon,\gamma)] = 0$, and

$$\mathbb{E}^{F^A_\varepsilon}[k(U,\theta^A_\varepsilon,\gamma)] \le v^A + \varepsilon\,.$$

Then $\underline{F}^A_\varepsilon$ solves the feasibility program (61) with $\theta = \theta^A_\varepsilon$ and $\kappa = \kappa^A_\varepsilon := \mathbb{E}^{F^A_\varepsilon}[k(U,\theta^A_\varepsilon,\gamma)]$. Note that $\mathbb{E}^{F^A_\varepsilon}[g_{4e}(U,\theta^A_\varepsilon,\gamma)] = 0$ also holds by construction. Therefore, $(\kappa^A_\varepsilon, \theta^A_\varepsilon)$ are feasible for the augmented form of Program B. It follows that $v^B \le \kappa^A_\varepsilon \le v^A + \varepsilon$ holds for each $\varepsilon > 0$. As $\varepsilon > 0$ is arbitrary, we have $v^B \le v^A$.

A similar argument applies when $v^A = -\infty$: for any $n \in \mathbb{N}$, we may choose $\theta^A_n \in \Theta$ and $F^A_n \in \mathcal{N}_\delta$ such that the constraints in Program A are satisfied and $\mathbb{E}^{F^A_n}[k(U,\theta^A_n,\gamma)] \le -n$. It follows that $v^B \leq -n$. As this is true for all $n \in \mathbb{N}$, we have $v^B = v^A$. ∎

Proof of Proposition 3.2. We prove the result for $\underline{F}_{\delta,\theta}$, the result for $\overline{F}_{\delta,\theta}$ follows similarly. We drop dependence of $\underline{F}$, $\underline{m}$, k, and g on (θ, γ) to simplify notation in what follows.

First, consider the case in which k depends on u. The dual formulation is justified by Proposition 2.1, replacing g by g_s in the dual formulation and the moment conditions (1a)–(1d) by (1a)–(1c) and $\mathbb{E}^F[g_{4s}(U, \theta, \gamma)] = 0$ in the statement of the Proposition. Note the primal and dual values are equal and finite and a dual solution exists (by Lemmas E.2 and E.5).

Differentiability of the objective function in (η, ζ, λ) is guaranteed by Assumption Φ. Also note that Assumption Φ(i) ensures $\dot{\phi}^\star \geq 0$. The first-order condition (FOC) for $\underline{\zeta}$ is

$$0 = \mathbb{E}^{F_*}\left[\dot{\phi}^\star(-\underline{\eta}^{-1}(k(U) + \underline{\zeta} + \underline{\lambda}' g_s(U)))\right] - 1$$

which implies $\mathbb{E}^{F_*}[\underline{m}_{\delta,\theta}] = 1$ and hence that $\underline{F}_{\delta,\theta}$ is a probability measure. The FOC for $\underline{\lambda}$ is

$$\begin{aligned}
0 &\geq \mathbb{E}^{F_*}\left[\dot{\phi}^\star(-\underline{\eta}^{-1}(k(U) + \underline{\zeta} + \underline{\lambda}' g_s(U)))g_1(U)\right] - P_1\,,\\
0 &= \mathbb{E}^{F_*}\left[\dot{\phi}^\star(-\underline{\eta}^{-1}(k(U) + \underline{\zeta} + \underline{\lambda}' g_s(U)))g_2(U)\right] - P_2\,,\\
0 &\geq \mathbb{E}^{F_*}\left[\dot{\phi}^\star(-\underline{\eta}^{-1}(k(U) + \underline{\zeta} + \underline{\lambda}' g_s(U)))g_3(U)\right]\,,\\
0 &= \mathbb{E}^{F_*}\left[\dot{\phi}^\star(-\underline{\eta}^{-1}(k(U) + \underline{\zeta} + \underline{\lambda}' g_s(U)))g_{4s}(U)\right]\,,
\end{aligned}$$

hence (1a)–(1c) and $\mathbb{E}^F[g_{4s}(U, \theta, \gamma)] = 0$ hold at (θ, γ, P) under $\underline{F}_{\delta,\theta}$. The FOC for $\underline{\eta} > 0$ is

$$\begin{aligned}
0 = \mathbb{E}^{F_*}&\left[\dot{\phi}^\star(-\underline{\eta}^{-1}(k(U) + \underline{\zeta} + \underline{\lambda}' g_s(U)))(-\underline{\eta}^{-1}(k(U) + \underline{\zeta} + \underline{\lambda}' g_s(U)))\right]\\
&- \mathbb{E}^{F_*}\left[\phi^\star(-\underline{\eta}^{-1}(k(U) + \underline{\zeta} + \underline{\lambda}' g_s(U)))\right] - \delta\,.
\end{aligned}$$

By Assumption Φ(i), we may write the convex conjugate $\phi^{\star\star}$ of $\phi^\star$ using its Legendre transform:

$$\phi^{\star\star}(x^\star) = x^\star(\dot{\phi}^\star)^{-1}(x^\star) - \phi^\star((\dot{\phi}^\star)^{-1}(x^\star))$$

for any $x^\star$ in the range of $\dot{\phi}^\star$ (Rockafellar, 1970, Theorem 26.4). Setting $x^\star = \dot{\phi}^\star(x)$ and noting that $\phi^{\star\star} = \phi$ holds by the Fenchel–Moreau theorem, we obtain

$$\phi(\dot{\phi}^\star(x)) = x\dot{\phi}^\star(x) - \phi^\star(x)\,.$$

It follows that we may rewrite the FOC for $\underline{\eta}$ as $\delta = \mathbb{E}^{F^*}\left[\phi(\underline{m}_{\delta,\theta}(U))\right]$ and so $\underline{F}_{\delta,\theta} \in \mathcal{N}_\delta$.

Now consider the case in which k does not depend on u. Lemma F.3 justifies the dual representation of the program (17), equality of the primal and dual values, and existence of a dual solution. A similar argument to the previous case shows that $\mathbb{E}^{F_*}[\underline{m}_{\delta,\theta}] = 1$ and hence that $\underline{F}_{\delta,\theta}$ is a probability measure, and that (1a)–(1c) and $\mathbb{E}^F[g_{4s}(U,\theta,\gamma)] = 0$ hold at (θ,γ,P) under $\underline{F}_{\delta,\theta}$. Finally, as there exists a distribution F with $D(F\|F_*) < \delta$ under which the moment conditions (1a)–(1c) and $\mathbb{E}^F[g_{4s}(U,\theta,\gamma)] = 0$ hold at (θ,γ,P). By construction, $D(\underline{F}_{\delta,\theta}\|F_*) \leq D(F\|F_*)$ for said F, and so $D(\underline{F}_{\delta,\theta}\|F_*) \leq \delta$, as required. ■

E.4 Proofs for Section 4

Proof of Proposition 4.1. The result is a consequence of Theorem 1 of Sason and Verdú (2016), which implies $D_{\phi_1}(F\|F_*) \leq \bar{a} D_{\phi_2}(F\|F_*)$. The inclusion $\mathcal{N}_{\phi_2,\delta} \subseteq \mathcal{N}_{\phi_1,\bar{a}\delta}$ for each $\delta > 0$ follows immediately. The result now follows from this inclusion, noting that Assumption Φ(ii) holding for $\mathcal{E}$ corresponding to ϕ_1 implies the values $\underline{\kappa}_{\phi_1,\bar{a}\delta}$ and $\underline{\kappa}_{\phi_1,\bar{a}\delta}$ are both finite. ■

E.5 Proofs for Section 6

We first present two lemmas which we shall use multiple times in the following proofs.

Lemma E.7 *Let Assumptions Φ and M(i)(v) hold and let $\{(F_n,\theta_n,\gamma_n,P_n)\}_{n\geq 1} \subseteq \mathcal{N}_\delta \times \Theta \times \Gamma \times \mathcal{P}$ with $(\gamma_n,P_n) \to (\tilde{\gamma},\tilde{P}) \in \Gamma \times \mathcal{P}$ and for which (1) holds under F_n at (θ_n,γ_n,P_n). Then: there exists a convergent subsequence $(F_{n_l},\theta_{n_l},\gamma_{n_l},P_{n_l}) \to (\tilde{F},\tilde{\theta},\tilde{\gamma},\tilde{P}) \in \mathcal{N}_\delta \times \Theta \times \Gamma \times \mathcal{P}$ along which (i) $\lim_{l\to\infty} \mathbb{E}^{F_{n_l}}[k(U,\theta_{n_l},\gamma_{n_l})] = \mathbb{E}^{\tilde{F}}[k(U,\tilde{\theta},\tilde{\gamma})]$ and similarly for each entry of $g_1,\ldots,g_4$, and (ii) the moment conditions (1) hold under $\tilde{F}$ at $(\tilde{\theta},\tilde{\gamma},\tilde{P})$.*

Proof of Lemma E.7. Let $m_n = \frac{\mathrm{d}F_n}{\mathrm{d}F_*}$. By Assumption M(v), $\{\theta_n\}_{n\geq 1}$ has a convergent subsequence $\{\theta_{n_l}\}_{l\geq 1}$. As $\{m_{n_l}\}_{l\geq 1}$ is $\|\cdot\|_\phi$-norm bounded (cf. Lemma D.2), taking a further subsequence if necessary we may assume $\{m_{n_l}\}_{l\geq 1}$ is $\mathcal{E}$-weakly convergent to $\tilde{m} \in \mathcal{L}$. By Lemma D.1 we have $\delta \geq \liminf_{l\to\infty} \mathbb{E}^{F_*}[\phi(m_{n_l}(U))] \geq \mathbb{E}^{F_*}[\phi(\tilde{m}(U))]$. By the triangle and Hölder inequalities,

$$\begin{aligned}&\left|\mathbb{E}^{F_{n_l}}[m_{n_l}(U)k(U,\theta_{n_l},\gamma_{n_l})] - \mathbb{E}^{F_*}[\tilde{m}(U)k(U,\tilde{\theta},\tilde{\gamma})]\right| \\ &\quad \leq |\mathbb{E}^{F_*}[(m_{n_l}(U)-\tilde{m}(U))k(U,\tilde{\theta},\tilde{\gamma})]| + \|m_{n_l}\|_\phi \|k(\cdot,\theta_{n_l},\gamma_{n_l}) - k(\cdot,\tilde{\theta},\tilde{\gamma})\|_\psi \to 0\end{aligned}$$

by $\mathcal{E}$-weak convergence and Assumption M(i). By similar arguments, we may deduce

$$\begin{aligned}\mathbb{E}^{F_*}[\tilde{m}(U)] = 1\,, \qquad & \mathbb{E}^{F_*}[\tilde{m}(U)g_1(U,\tilde{\theta},\tilde{\gamma})] \leq \tilde{P}_1\,, \qquad && \mathbb{E}^{F_*}[\tilde{m}(U)g_2(U,\tilde{\theta},\tilde{\gamma})] = \tilde{P}_2\,, \\ & \mathbb{E}^{F_*}[\tilde{m}(U)g_3(U,\tilde{\theta},\tilde{\gamma})] \leq 0\,, && \mathbb{E}^{F_*}[\tilde{m}(U)g_4(U,\tilde{\theta},\tilde{\gamma})] = 0\end{aligned}$$

all hold. ■

Lemma E.8 *Let Assumptions Φ and M(i)(ii) hold, let Condition S' hold at (θ,γ_0,P_0), and let $\Delta^\star(\theta;\gamma_0,P_0) < \delta$. Then: there is a neighborhood N of (θ,γ_0,P_0) such that for all $(\tilde{\theta},\tilde{\gamma},\tilde{P}) \in N$*

we have that Condition S' holds at $(\tilde{\theta}, \tilde{\gamma}, \tilde{P})$*,* $\Delta^\star(\tilde{\theta}; \tilde{\gamma}, \tilde{P}) < \delta$*,* $\underline{K}_\delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P}) = \underline{K}^\star_\delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P})$ *and* $\overline{K}_\delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P}) = \overline{K}^\star_\delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P})$*.*

Proof of Lemma E.8. By Lemma F.2 there is a neighborhood N' of (θ, γ_0, P_0) such that Condition S' holds at $(\tilde{\theta}, \tilde{\gamma}, \tilde{P})$ for all $(\tilde{\theta}, \tilde{\gamma}, \tilde{P}) \in N'$. Moreover, the inequality $\Delta^\star(\tilde{\theta}, \tilde{\gamma}, \tilde{P}) < \delta$ holds on a neighborhood N'' of (θ, γ, P) by continuity of $\Delta^\star$ at (θ, γ_0, P_0) (cf. Lemma F.4). It follows by Lemma F.3 that for each $(\tilde{\theta}, \tilde{\gamma}, \tilde{P}) \in N' \cap N''$ there exists F with $D_\phi(F \| F_*) < \delta$ satisfying (1) at $(\tilde{\theta}, \tilde{\gamma}, \tilde{P})$. Therefore, $\underline{K}_\delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P}) = \underline{K}^\star_\delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P})$ on $N' \cap N''$ by Lemma E.5. ■

Lemma E.9 *Let Assumptions* Φ *and M(i)(iii)(iv)(v) hold. Then* $\underline{\kappa}_\delta$ *and* $\overline{\kappa}_\delta$ *are finite, and:*

$$\underline{\kappa}_\delta = \inf_{\theta \in \Theta_\delta(\gamma_0, P_0)} \underline{K}_\delta(\theta; \gamma_0, P_0) \qquad\qquad \overline{\kappa}_\delta = \sup_{\theta \in \Theta_\delta(\gamma_0, P_0)} \overline{K}_\delta(\theta; \gamma_0, P_0) \,.$$

Proof of Lemma E.9. We prove the result only for $\underline{\kappa}_\delta$; the result for $\overline{\kappa}_\delta$ follows similarly.

Finiteness of $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ follows by Assumptions Φ and M(i)(v) and Hölder's inequality. To simplify notation, we suppress dependence of $\Theta_\delta(\gamma_0, P_0)$ on (γ_0, P_0) in what follows. Suppose there is $\underline{\theta} \not\in \Theta_\delta$ with $\underline{K}_\delta(\underline{\theta}; \gamma_0, P_0) < \inf_{\theta \in \Theta_\delta} \underline{K}_\delta(\theta; \gamma_0, P_0)$. Then there must exist $F_{\underline{\theta}} \in \mathcal{N}_\delta$ satisfying the moment conditions at $(\underline{\theta}, \gamma_0, P_0)$. As $\Delta^\star(\underline{\theta}; \gamma_0, P_0) = \delta$, it follows by convexity of ϕ that $F_{\underline{\theta}}$ must be the unique such F. Therefore

$$\mathbb{E}^{F_{\underline{\theta}}}[k(U, \underline{\theta}, \gamma_0)] = \underline{K}_\delta(\underline{\theta}; \gamma_0, P_0) < \inf_{\theta \in \Theta_\delta} \underline{K}_\delta(\theta; \gamma_0, P_0) \le \inf_{\theta \in \Theta_\delta} \mathbb{E}^{F_\theta}[k(U, \theta_0, \gamma_0)] \,, \tag{62}$$

where, for each $\theta \in \Theta_\delta$, the distribution F_θ solves $\inf_F D_\phi(F \| F_*)$ subject to (1). Existence of such an F_θ follows by similar arguments to the proof of Lemma E.2(ii) and its uniqueness follows by strict convexity of ϕ.

Choose $\{\theta_n\}_{n \ge 1} \subset \Theta_\delta$ with $\theta_n \to \underline{\theta}$ (we may choose such a sequence by Assumption M(iv)). By Lemma E.7, there is a subsequence $\{(\theta_{n_l}, F_{\theta_{n_l}})\}_{l \ge 1}$ with $(\theta_{n_l}, F_{\theta_{n_l}}) \to (\underline{\theta}, \underline{F})$ for some $\underline{F} \in \mathcal{N}_\delta$ for which (1) holds under $\underline{F}$ at $(\underline{\theta}, \gamma_0, P_0)$. It follows by uniqueness of $F_{\underline{\theta}}$ established above that $\underline{F} = F_{\underline{\theta}}$. By Lemma E.7, we therefore have

$$\inf_{\theta \in \Theta_\delta} \mathbb{E}^{F_\theta}[k(U, \theta, \gamma_0)] \le \lim_{l \to \infty} \mathbb{E}^{F_{\theta_{n_l}}}[k(U, \theta_{n_l}, \gamma_0)] = \mathbb{E}^{F_{\underline{\theta}}}[k(U, \underline{\theta}, \gamma_0)] \,,$$

which contradicts (62). ■

Define

$$\underline{b}_\delta(\gamma, P) = \inf_{\theta \in \Theta_\delta(\gamma, P)} \underline{K}_\delta(\theta; \gamma, P) \,, \quad \overline{b}_\delta(\gamma, P) = \inf_{\theta \in \Theta_\delta(\gamma, P)} \overline{K}_\delta(\theta; \gamma, P) \,.$$

Lemma E.10 *Let Assumptions* Φ *and M(i)–(v) hold. Then:* $\underline{b}_\delta(\gamma, P)$ *and* $\overline{b}_\delta(\gamma, P)$ *are continuous at* (γ_0, P_0)*.*

Proof of Lemma E.10. We prove the result only for $\underline{b}_\delta$; the result for $\overline{b}_\delta$ follows similarly.

Fix $\varepsilon > 0$. By Lemma E.9, we may choose $\theta_\varepsilon \in \Theta_\delta(\gamma_0, P_0)$ such that $\underline{K}_\delta(\theta_\varepsilon; \gamma_0, P_0) < \underline{b}_\delta(\gamma_0, P_0) + \varepsilon$. By Lemma E.8 and Assumption M(ii)(iii), $\underline{K}_\delta(\theta_\varepsilon; \gamma, P) = \underline{K}^\star_\delta(\theta_\varepsilon; \gamma, P)$ and $\Delta^\star(\theta_\varepsilon; \gamma, P) < \delta$ both hold for all (γ, P) in a neighborhood N of (γ_0, P_0). Moreover, Lemma F.5 implies $\underline{K}^\star_\delta(\theta_\varepsilon; \gamma, P)) < \underline{K}^\star_\delta(\theta_\varepsilon; \gamma_0, P_0) + \varepsilon$ for all (γ, P) in a neighborhood N' of (γ_0, P_0). Then on $N \cap N'$,

$$\underline{b}_\delta(\gamma, P) \leq \underline{K}_\delta(\theta_\varepsilon; \gamma, P) = \underline{K}^\star_\delta(\theta_\varepsilon; \gamma, P) < \underline{K}^\star_\delta(\theta_\varepsilon; \gamma_0, P_0) + \varepsilon = \underline{K}_\delta(\theta_\varepsilon; \gamma_0, P_0) + \varepsilon < \underline{b}_\delta(\gamma_0, P_0) + 2\varepsilon\,,$$

proving $\underline{b}_\delta(\gamma, P)$ is u.s.c. at (γ_0, P_0).

To establish l.s.c., suppose there is $\varepsilon > 0$ and a sequence $\{(\gamma_n, P_n)\}_{n \geq 1}$ converging to (γ_0, P_0) along which

$$\underline{b}_\delta(\gamma_n, P_n) \leq \underline{b}_\delta(\gamma_0, P_0) - 2\varepsilon\,. \tag{63}$$

The set $\Theta_\delta(\gamma_n, P_n)$ is nonempty for n sufficiently large by Lemma F.4 and Assumptions M(ii)(iii). For each n sufficiently large, choose $\theta_n \in \Theta_\delta(\gamma_n, P_n)$ and $F_n \in \mathcal{N}_\delta$ for which

$$\mathbb{E}^{F_n}[k(U, \theta_n, \gamma_n)] < \underline{b}_\delta(\gamma_n, P_n) + \varepsilon\,. \tag{64}$$

By Lemma E.7 there is $\{(F_{n_l}, \theta_{n_l}, \gamma_{n_l}, P_{n_l})\}_{l \geq 1}$ with $(F_{n_l}, \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \to (\underline{F}, \underline{\theta}, \gamma_0, P_0)$ for some $\underline{F} \in \mathcal{N}_\delta$ and $\underline{\theta} \in \Theta$, such that (1) holds under $\underline{F}$ at $(\underline{\theta}, \gamma_0, P_0)$, and for which

$$\lim_{l \to \infty} \mathbb{E}^{F_{n_l}}[k(U, \theta_{n_l}, \gamma_{n_l})] = \mathbb{E}^{\underline{F}}[k(U, \underline{\theta}, \gamma_0)] \geq \underline{K}_\delta(\underline{\theta}; \gamma_0, P_0)\,.$$

In view of (63) and (64) and Lemma E.9, this implies $\underline{K}_\delta(\underline{\theta}; \gamma_0, P_0) \leq \underline{b}_\delta(\gamma_0, P_0) - \varepsilon = \underline{\kappa}_\delta - \varepsilon$, contradicting the definition of $\underline{\kappa}_\delta$. ■

Proof of Theorem 6.1. Immediate from Lemma E.10 and Slutsky's theorem, noting that $\hat{\underline{\kappa}}_\delta = \underline{b}_\delta(\hat{\gamma}, \hat{P})$ and $\hat{\overline{\kappa}}_\delta = \overline{b}_\delta(\hat{\gamma}, \hat{P})$. ■

In the remainder of this subsection we drop dependence of all quantities on γ.

Proof of Theorem 6.2. We prove the result only for $\underline{b}_\delta$; the result for $\overline{b}_\delta$ follows similarly.

Step 1: We first show $\underline{\Theta}_\delta(P_0)$ is nonempty and compact. For nonemptiness, choose $\{\theta_n\}_{n \geq 1}$ such that $\underline{K}_\delta(\theta_n; P_0) \downarrow \underline{\kappa}_\delta$. Let F_n solve the primal problem for θ_n. By Lemma E.7, there is a subsequence $(F_{n_l}, \theta_{n_l}) \to (\underline{F}, \underline{\theta})$ with $\underline{F} \in \mathcal{N}_\delta$ and $\theta \in \Theta$ such that (1) holds under $\underline{F}$ at $(\underline{\theta}, P_0)$ and for which

$$\underline{\kappa}_\delta = \lim_{l \to \infty} \mathbb{E}^{F_{n_l}}[k(U, \theta_{n_l})] = \mathbb{E}^{\underline{F}}[k(U, \underline{\theta})]\,.$$

Therefore, $\Theta_\delta(P_0)$ is nonempty. The proof that $\underline{\Theta}_\delta(P_0)$ is closed follows by similar arguments to the proof of nonemptiness. Compactness now follows by Assumption M(v).

Step 2: We now prove directional differentiability. Let $P_n = P_0 + t_n h_n$ with $t_n \downarrow t$ and

$h_n \to h \in \mathbb{R}^{d_1+d_2}$. Choose any $\underline{\theta} \in \underline{\Theta}_\delta(P_0)$. By Lemma E.8 (using Assumptions M(iii)(vi)),

$$\underline{b}_\delta(P_n) - \underline{b}_\delta(P_0) \leq \underline{K}_\delta(\underline{\theta}; P_n) - \underline{K}_\delta(\underline{\theta}; P_0) = \underline{K}^\star_\delta(\underline{\theta}; P_n) - \underline{K}^\star_\delta(\underline{\theta}; P_0) \leq t_n \times -\underline{\lambda}'_{12} h_n$$

for all n sufficiently large, where the final inequality holds for any $\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}; P_n)$. By Lemma F.5, compactness of $\underline{\Lambda}_\delta(\underline{\theta}; P_0)$ (cf. Lemma E.5 using Assumptions M(iii)(vi)), and the above chain of inequalities, we obtain

$$\limsup_{n\to\infty} \frac{\underline{b}_\delta(P_n) - \underline{b}_\delta(P_0)}{t_n} \leq \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}; P_0)} -\underline{\lambda}'_{12} h \,.$$

As this inequality holds for any $\underline{\theta} \in \underline{\Theta}_\delta$, taking the infimum over $\underline{\theta} \in \underline{\Theta}_\delta$ yields

$$\limsup_{n\to\infty} \frac{\underline{b}_\delta(P_n) - \underline{b}_\delta(P_0)}{t_n} \leq \inf_{\theta \in \underline{\Theta}_\delta} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}; P_0)} -\underline{\lambda}'_{12} h \,. \tag{65}$$

For the lower bound, choose $\{\theta_n\}_{n\geq 1}$ with $\theta_n \in \Theta_\delta(P_n)$ for all n sufficiently large and for which $\underline{K}_\delta(\theta_n; P_n) \leq \underline{b}_\delta(P_n) + t_n^2$. Take a subsequence $\{\theta_{n_l}\}_{l\geq 1}$. By Assumption M(v), we may extract a convergent subsequence $\theta_{n_{l_j}} \to \underline{\theta} \in \Theta$. By similar arguments to Step 1 we may deduce $\underline{\theta} \in \underline{\Theta}_\delta$. By Lemma E.8 (using Assumptions M(iii)(vi)), for j sufficiently large

$$\begin{aligned}\underline{b}_\delta(P_{n_{l_j}}) - \underline{b}_\delta(P_0) &\geq \underline{K}_\delta(\theta_{n_{l_j}}; P_{n_{l_j}}) - \underline{K}_\delta(\theta_{n_{l_j}}; P_0) - t^2_{n_{l_j}} \\ &= \underline{K}^\star_\delta(\theta_{n_{l_j}}; P_{n_{l_j}}) - \underline{K}^\star_\delta(\theta_{n_{l_j}}; P_0) - t^2_{n_{l_j}} \geq t_{n_{l_j}} \times -\underline{\lambda}'_{12} h_{n_{l_j}} - t^2_{n_{l_j}} \,,\end{aligned}$$

where the final inequality holds for any $\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta_{n_{l_j}}; P_0)$. By Assumption M(vii), we may choose $\{\underline{\lambda}_{12,n_{l_j}}\}_{j\geq 1}$ with $\underline{\lambda}_{12,n_{l_j}} \in \Lambda_\delta(\theta_{n_{l_j}}; P_0)$ for each j, for which

$$-\underline{\lambda}'_{12,n_{l_j}} h \to \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}; P_0)} -\underline{\lambda}'_{12} h \,.$$

Therefore,

$$\liminf_{j\to\infty} \frac{\underline{b}_\delta(P_{n_{l_j}}) - \underline{b}_\delta(P_0)}{t_{n_{l_j}}} \geq \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}; P_0)} -\underline{\lambda}'_{12} h \geq \inf_{\theta \in \underline{\Theta}_\delta} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta; P_0)} -\underline{\lambda}'_{12} h \,.$$

As the lower bound does not depend on the subsequence $\{\theta_{n_l}\}_{l\geq 1}$, we have

$$\liminf_{n\to\infty} \frac{\underline{b}_\delta(P_n) - \underline{b}_\delta(P_0)}{t_n} \geq \inf_{\theta \in \underline{\Theta}_\delta} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta; P_0)} -\underline{\lambda}'_{12} h \,, \tag{66}$$

proving directional differentiability. Finally, note that under Assumption M(vii) the function $\theta \mapsto \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta; P_0)} -\underline{\lambda}'_{12} h$ is continuous on $\underline{\Theta}_\delta$ (rather than just u.s.c., as implied by Lemma F.5). It follows by Step 1 that that the infima in (65) and (66) can be replaced by minima.

Step 3: We now derive the joint asymptotic distribution. In view of Steps 2 and 3, this result follows from Theorem 2.1 of Shapiro (1991) and $\sqrt{n}(\hat{P}-P) \to_d N(0,\Sigma)$. ■

Before proving Theorem 6.3, we first present a preliminary result which is useful for verifying some conditions in Fang and Santos (2019).

Lemma E.11 *Let the conditions of Theorem 6.3 hold. Then:*

(i) $|\widehat{d\underline{b}}_{\delta,P_0}[h_1] - \widehat{d\underline{b}}_{\delta,P_0}[h_2]| \le \underline{C}_n \|h_1 - h_2\|$ *and* $|\widehat{d\overline{b}}_{\delta,P_0}[h_1] - \widehat{d\overline{b}}_{\delta,P_0}[h_2]| \le \overline{C}_n \|h_1 - h_2\|$ *for all* $h_1, h_2 \in \mathbb{R}^{d_1+d_2}$ *with* $\underline{C}_n = O_p(1)$ *and* $\overline{C}_n = O_p(1)$*;*

(ii) $\widehat{d\underline{b}}_{\delta,P_0}[h] \to_p d\underline{b}_{\delta,P_0}[h]$ *and* $\widehat{d\overline{b}}_{\delta,P_0}[h] \to_p d\overline{b}_{\delta,P_0}[h]$ *for all* $h \in \mathbb{R}^{d_1+d_2}$.

Proof of Lemma E.11. We prove the results for the lower values; the results for the upper values follow similarly. Let $\underline{\Theta}_{\delta,a}(P) = \{\theta \in \Theta_\delta(P) : \underline{K}_\delta(\theta;P) \le \underline{b}_\delta(P) + a\}$.

Step 1: We first show there exists $\{a_n\}_{n\ge 1}$ with $a_n \downarrow 0$ such that $\underline{\Theta}_\delta \subseteq \hat{\underline{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a_n}(P_0)$ holds with probability approaching one (wpa1).

To do so, we first show that there is a neighborhood N of P_0 such that for any $P \in N$, for all $\theta \in \underline{\Theta}_\delta$ we have $\underline{K}_\delta(\theta;P) = \underline{K}^\star_\delta(\theta;P)$, $\Delta^\star(\theta;P) < \delta$, and Condition S' holds at (θ,P). We prove this claim by contradiction. Suppose we can choose $P_n \to P_0$ and $\{\theta_n\}_{n\ge1} \subseteq \underline{\Theta}_\delta$ such that Condition S' doesn't hold at (θ_n, P_n) and/or $\Delta^\star(\theta_n;P_n) \ge \delta$ and/or $\underline{K}_\delta(\theta_n;P_n) \neq \underline{K}^\star_\delta(\theta_n;P_n)$ for each n. In view of Step 1 of the proof of Theorem 6.2, we can extract a convergent subsequence $\{(\theta_{n_l}, P_{n_l})\}_{l\ge1}$ with $\theta_{n_l} \to \underline{\theta} \in \underline{\Theta}_\delta$. Then by Lemma E.8 we must have that Condition S' holds at (θ_{n_l}, P_{n_l}) and $\Delta^\star(\theta_{n_l};P_{n_l}) < \delta$ and $\underline{K}_\delta(\theta_{n_l};P_{n_l}) = \underline{K}^\star_\delta(\theta_{n_l};P_{n_l})$ all hold for all l sufficiently large, a contradiction.

This intermediate result and consistency of $\hat{P}$ implies that $\underline{\Theta}_\delta \subseteq \Theta_\delta(\hat{P})$ wpa1. It also implies that $\Lambda_\delta(\theta;P)$ is nonempty and compact for all $\theta \in \underline{\Theta}_\delta$ for any $P \in N$ (cf. Lemma E.5).

By similar arguments to the proof of Theorem 6.2, for any $P \in N$ we have

$$\sup_{\theta \in \underline{\Theta}_\delta} \underline{K}_\delta(\theta;P) - \underline{b}_\delta(P) \le \sup_{\theta \in \underline{\Theta}_\delta} \max_{\underline{\lambda}_{12} \in \Lambda_\delta(\theta;P)} -\underline{\lambda}_{12}'(P-P_0) + \underline{b}_\delta(P_0) - \underline{b}_\delta(P)\,.$$

As $\underline{\Theta}_\delta$ is compact and $(\theta,P) \mapsto \max_{\underline{\lambda}_{12}\in\Lambda_\delta(\theta;P)} \|\underline{\lambda}_{12}\|$ is u.s.c. (by Lemma F.5) on $\underline{\Theta}_\delta \times N$, $\sup_{\theta\in\underline{\Theta}_\delta} \max_{\underline{\lambda}_{12}\in\Lambda_\delta(\theta;P)} \|\underline{\lambda}_{12}'\| \le C$ holds on a neighborhood N' of P_0 for some $C < \infty$. It follows by Theorem 6.2 that $\sup_{\theta\in\underline{\Theta}_\delta} K_\delta(\theta;\hat{P}) - \underline{b}_\delta(\hat{P}) \le O_p(\sqrt{n})$ and so $\underline{\Theta}_\delta \subseteq \hat{\underline{\Theta}}_{\delta,n}$ wpa1.

We now prove the remaining inclusion. By the almost sure representation theorem (Shapiro, 1991, Theorem A1), there exists a sequence of random vectors $\{(Z_n, \hat{\nu}_n)\}_{n\ge1}$ and a random vector Z defined on a single probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $Z_n =_d \sqrt{n}(\hat{P} - P_0)$, $\hat{\nu}_n =_d \hat{\nu}$, $Z \sim N(0,\Sigma)$, and $(Z_n, \hat{\nu}_n) \to_{a.s.} (Z,\nu)$. Let $P_n = P_0 + n^{-1/2} Z_n$ so that $P_n =_d \hat{P}$. Fix any $\omega \in \Omega$ for which $(Z_n(\omega), \hat{\nu}_n(\omega)) \to (Z(\omega), \nu(\omega))$.

We wish to show that for any fixed $a > 0$, the inclusion $\hat{\underline{\Theta}}_{\delta,n}(\omega) := \{\theta \in \Theta_\delta(P_n(\omega)) : \underline{K}_\delta(\theta; P_n(\omega)) \le \underline{b}_\delta(P_n(\omega)) + \nu(\omega)\sqrt{\log n/n}\} \subseteq \underline{\Theta}_{\delta,a}(P_0)$ holds for all n sufficiently large. To do so, we proceed by contradiction. By Lemma E.8, $\Theta_\delta(P_n(\omega))$ and hence $\underline{\hat{\Theta}}_{\delta,n}(\omega)$ are nonempty for n sufficiently large. Suppose there is $\{\theta_{n_l}(\omega)\}_{l\geq 1}$ with $\theta_{n_l}(\omega) \in \underline{\hat{\Theta}}_{\delta,n}(\omega) \setminus \underline{\Theta}_{\delta,a}(P_0)$ for all l. By Assumption M(v) (taking a subsequence if necessary) we may assume $\theta_{n_l}(\omega) \to \underline{\theta}(\omega) \in \Theta$. By similar arguments to Step 1 in the proof of Theorem 6.2, we may deduce $\underline{\theta}(\omega) \in \underline{\Theta}_\delta$. Then by Lemmas E.8 and F.5, we have $\lim_{l\to\infty} \underline{K}_\delta(\theta_{n_l}(\omega); P_0) = \lim_{l\to\infty} \underline{K}^\star_\delta(\theta_{n_l}(\omega); P_0) = \underline{K}^\star_\delta(\underline{\theta}(\omega); P_0) = \underline{K}_\delta(\underline{\theta}(\omega); P_0) = \underline{b}_\delta(P_0)$ and $\lim_{l\to\infty} \Delta^\star(\theta_{n_l}(\omega); P_0) < \delta$. Therefore, $\theta_{n_l}(\omega) \in \underline{\Theta}_{\delta,a}(P_0)$ for l sufficiently large, a contradiction.

As $\underline{\hat{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a}(P_0)$ holds wpa1 for any fixed $a > 0$, we may therefore choose $\{a_n\}_{n\geq 1}$ with $a_n \downarrow 0$ such that $\underline{\hat{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a_n}(P_0)$ holds wpa1.

Step 2: We prove part (i). First note that we may choose sufficiently small $a > 0$ such that both Condition S' holds at (θ, P_0) and $\Delta^\star(\theta; P_0) < \delta$ for all $\theta \in \underline{\Theta}_{\delta,a}(P_0)$. If not, we may choose $\{a_n\}_{n\geq 1}$ with $a_n \downarrow 0$ and $\theta_n \in \underline{\Theta}_{\delta,a_n}(P_0)$ for which Condition S' does not hold and/or $\Delta^\star(\theta_n; P_0) \geq \delta$. By Assumption M(v), we can extract a convergent subsequence $\{(\theta_{n_l}, P_{n_l})\}_{l\geq 1}$ with $\theta_{n_l} \to \underline{\theta}$. By similar arguments to Step 1 of the proof of Theorem 6.2, we may deduce in fact that $\underline{\theta} \in \underline{\Theta}_\delta$. Then by Lemma E.8 we must have that Condition S' holds at (θ_{n_l}, P_0) and $\Delta^\star(\theta_{n_l}; P_0) < \delta$ for all l sufficiently large, a contradiction.

By similar arguments to Step 1, we may choose a neighborhood N of P_0 such that for any $P \in N$ both Condition S' holds at (θ, P) and $\Delta^\star(\theta; P) < \delta$ for all $\theta \in \underline{\Theta}_{\delta,a}(P_0)$. Therefore, $\Lambda_\delta(\theta; P)$ is compact and nonempty for all $(\theta, P) \in \underline{\Theta}_{\delta,a}(P_0) \times N$. As $\underline{\Theta}_{\delta,a}(P_0)$ is compact (which follows by similar arguments to Step 1 in the proof of Theorem 6.2), and $(\theta, P) \mapsto \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta;P)} \|\underline{\lambda}_{12}\|$ is u.s.c. (by Lemma F.5) on $\underline{\Theta}_{\delta,a}(P_0) \times N$, we may deduce that $\sup_{\theta\in\underline{\Theta}_{\delta,a}(P_0)} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta;P)} \|\underline{\lambda}'_{12}\| \leq C$ holds on a neighborhood N' of P_0 for some $C < \infty$. Now, as $\hat{P} \in N \cap N'$ holds wpa1 and $\underline{\hat{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a_n}(P_0)$ holds wpa1, it follows that

$$\left| \inf_{\theta\in\underline{\hat{\Theta}}_{\delta,n}} \sup_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\theta;\hat{P})} -\underline{\lambda}'_{12} h_1 - \inf_{\theta\in\underline{\hat{\Theta}}_{\delta,n}} \sup_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\theta;\hat{P})} -\underline{\lambda}'_{12} h_2 \right| \leq C\|h_1 - h_2\| \quad \text{for all } h_1, h_2 \in \mathbb{R}^{d_1+d_2}$$

holds wpa1, proving part (i).

Step 3: We prove part (ii). In view of Step 1, it suffices to show

$$\inf_{\theta\in\underline{\Theta}_\delta} \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\theta,\hat{P})} -\underline{\lambda}'_{12} h \to_p d\underline{b}_{\delta,P_0}[h]\,, \qquad \inf_{\theta\in\underline{\Theta}_{\delta,a_n}(P_0)} \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\theta,\hat{P})} -\underline{\lambda}'_{12} h \to_p d\underline{b}_{\delta,P_0}[h]\,.$$

Using the almost sure representation from Step 1, fix any $\omega \in \Omega$ for which $Z_n(\omega) \to Z(\omega)$. Let $\underline{\theta}_h$ solve $\min_{\theta\in\underline{\Theta}_\delta} \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\theta;P_0)} -\underline{\lambda}_{12} h$. By Lemma F.5, we have

$$\begin{aligned}
&\limsup_{n\to\infty} \inf_{\theta\in\underline{\Theta}_\delta} \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\theta,P_n(\omega))} -\underline{\lambda}'_{12} h \\
&\leq \limsup_{n\to\infty} \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\underline{\theta}_h,P_n(\omega))} -\underline{\lambda}'_{12} h \leq \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\underline{\theta}_h,P_0)} -\underline{\lambda}'_{12} h = d\underline{b}_{\delta,P_0}[h]\,. \qquad (67)
\end{aligned}$$

Now choose $\underline{\theta}_n(\omega) \in \underline{\Theta}_{\delta,a_n}(P_0)$ such that

$$\inf_{\theta \in \underline{\Theta}_{\delta,a_n}(P_0)} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta, P_n(\omega))} -\underline{\lambda}_{12}' h \geq \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta_n(\omega), P_n(\omega))} -\underline{\lambda}_{12}' h - \frac{1}{n} \,.$$

By Assumption M(v), for any subsequence $\{\theta_{n_l}(\omega), P_{n_l}(\omega)\}_{l \geq 1}$ we may extract a subsequence $\{\theta_{n_{l_j}}(\omega), P_{n_{l_j}}(\omega)\}_{j \geq 1}$ with $\theta_{n_{l_j}}(\omega) \to \underline{\theta}(\omega) \in \Theta$. By similar arguments to Step 1 of the proof of Theorem 6.2, we may deduce in fact that $\underline{\theta}(\omega) \in \underline{\Theta}_\delta$. By Lemma F.5 and lower hemicontinuity of $(\theta, P) \mapsto \underline{\Lambda}_\delta(\theta, P)$, writing $P_j = P_{n_{l_j}}(\omega)$, $\theta_j = \theta_{n_{l_j}}(\omega)$, and $\underline{\Theta}_{\delta,a_{n_{l_j}}}(P_0) = \underline{\Theta}_{\delta,a_j}(P_0)$, we have

$$\begin{aligned}
&\liminf_{j\to\infty} \inf_{\theta \in \underline{\Theta}_{\delta,a_j}(P_0)} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta, P_j)} -\underline{\lambda}_{12}' h \\
&\geq \liminf_{j\to\infty} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta_j, P_j)} -\underline{\lambda}_{12}' h \geq \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}(\omega), P_0)} -\underline{\lambda}_{12}' h \geq d\underline{b}_{\delta,P_0}[h] \,.
\end{aligned}$$

As the lower bound on the right-hand side does not depend on the subsequence chosen, we therefore have $\liminf_{n\to\infty} \inf_{\theta \in \underline{\Theta}_{\delta,a_n}(P_0)} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta, P_n(\omega))} -\underline{\lambda}_{12}' h \geq d\underline{b}_{\delta,P_0}[h]$. This, in conjunction with the upper bound (67), completes the proof. ■

Proof of Theorem 6.3. We verify the conditions of Theorem 3.2 of Fang and Santos (2019). Their Assumptions 1 and 2 holds by Theorem 6.2 and because $\sqrt{n}(\hat{P} - P_0) \to_d N(0, \Sigma)$ with Σ finite, respectively. Their Assumption 3 is assumed directly. Finally, Lemma E.11 shows that $\widehat{d\underline{b}}_{\delta,P_0}$ and $\widehat{d\overline{b}}_{\delta,P_0}$ satisfy the sufficient conditions for Assumption 4 of Fang and Santos (2019) presented in their Remark 3.4. Therefore, the proposed bootstrap procedure is consistent for the asymptotic distribution derived in Theorem 6.2. Correct coverage now follows by the continuity conditions on the distribution functions, which ensure the bootstrap estimates of the quantiles are consistent for the true quantiles of $\underline{G}_\delta$ and $\overline{G}_\delta$. The coverage of CS_α may be deduced by the Bonferroni inequality, leading to the inequality in the statement of the theorem. ■

Proof of Theorem 6.4. We prove the result only for $CS^{1-\alpha}_{\delta,L}$; the result for the other CSs follow similarly. Say that $P_0 \in CS^{1-\alpha}_{P_0}$ if $P_{10} \leq \hat{P}^{1-\alpha}_{1,U}$ and $P_{20} \in [\hat{P}^{1-\alpha}_{2,L}, \hat{P}^{1-\alpha}_{2,U}]$ both hold. By Lemma E.9, for each $\varepsilon > 0$ we may choose $\theta_\varepsilon \in \Theta_\delta(P_0)$ such that $\underline{K}_\delta(\theta_\varepsilon; P_0) < \underline{\kappa}_\delta + \varepsilon$. Let F_{θ_ε} solve the primal problem $\Delta(\theta_\varepsilon; P_0)$ (see Appendix F.3). Whenever $P_0 \in CS^{1-\alpha}_{P_0}$ holds, F_{θ_ε} also satisfies the "relaxed" moment conditions used for computing $\hat{\underline{\kappa}}_{\delta,1-\alpha}$, so it follows that $\Delta^\star(\theta_\varepsilon; \hat{P}_{1-\alpha}) < \delta$. Moreover, as the primal solution for $\underline{K}_\delta(\theta_\varepsilon; P_0)$ is feasible for the relaxed problem whenever $P_0 \in CS^{1-\alpha}_{P_0}$, by Lemma E.5 we have

$$\hat{\underline{\kappa}}_{\delta,1-\alpha} \leq \underline{K}^\star_{\delta,cs}(\theta_\varepsilon; \hat{P}_{1-\alpha}) = \underline{K}_{\delta,cs}(\theta_\varepsilon; \hat{P}_{1-\alpha}) \leq \underline{K}_\delta(\theta_\varepsilon; P_0) < \underline{\kappa}_\delta + \varepsilon \,,$$

and so $\underline{\kappa}_\delta \geq \hat{\underline{\kappa}}_{\delta,1-\alpha}$ holds whenever $P_0 \in CS^{1-\alpha}_{P_0}$. The desired coverage now follows by (29). ■

E.6 Proofs for Section 7

Proof of Theorem 7.1. First note that under the GMM-type regularity conditions, $k(\cdot,\theta,\gamma_0)$ and each entry of $g(\cdot,\theta,\gamma_0,P_{20})$ belong to $L^2(F_*)$ for all θ in a neighborhood of θ_*.

Step 1: We first prove a lower bound on s. To simplify notation, we drop dependence of g on (γ, P_2) and k on γ. Take any $b \in L^2(F_*)$ with $\mathbb{E}^{F_*}[b(U)] = 0$. Define $\mathbb{M} : L^2(F_*) \to L^2(F_*)$ by

$$\mathbb{M}b = b - \mathbb{E}^{F_*}[b(U)g_*(U)'](V^{-1} - V^{-1}G(G'V^{-1}G)^{-1}G'V^{-1})g_* \,.$$

Note $\mathbb{M}b = b$ if the model is just-identified. Using a standard construction (see, e.g. Example 3.2.1 in Bickel, Klaassen, Ritov, and Wellner (1993)), we define a smooth parametric family $\{F_t : t \in (-1,1)\}$ passing through F_* at $t = 0$ by

$$\frac{\mathrm{d}F_t}{\mathrm{d}F_*} = \frac{\upsilon(t\mathbb{M}b)}{\mathbb{E}^{F_*}[\upsilon(t\mathbb{M}b(U))]}\,, \quad \text{where } \upsilon(x) = \frac{2}{1+e^{-2x}}\,.$$

Fix any $d_\theta \times (d_2 + d_4)$ matrix A of full rank. Premultiplying g by A yields a just-identified system. By the implicit function theorem and invertibility of AG, there exists $\varepsilon > 0$ such that $\mathbb{E}^{F_t}[Ag(U,\theta)] = 0$ has a unique solution $\theta(F_t) \in \Theta$ for all $t \in (-\varepsilon, \varepsilon)$, and

$$\left.\frac{d\theta(F_t)}{dt}\right|_{t=0} = -(AG)^{-1}A\mathbb{E}^{F_*}[g_*(U)\mathbb{M}b(U)]\,.$$

Writing $\kappa(F_t) = \mathbb{E}^{F_t}[k(U,\theta(F_t))]$, we therefore have

$$\begin{aligned}
\left.\frac{d\kappa(F_t)}{dt}\right|_{t=0} &= \mathbb{E}^{F_*}[k_*(U)\mathbb{M}b(U)] - J'(AG)^{-1}A\mathbb{E}^{F_*}[g_*(U)\mathbb{M}b(U)] \\
&= \mathbb{E}^{F_*}[\tilde{\iota}(U)\mathbb{M}b(U)] \\
&= \mathbb{E}^{F_*}[\mathbb{M}\tilde{\iota}(U)\mathbb{M}b(U)]\,,
\end{aligned}$$

where $\tilde{\iota}(u) = k_*(u) - \kappa_* - J'(AG)^{-1}Ag_*(u)$ and the final line is because $\mathbb{M}$ is an orthogonal projection. However, note that for any A, we have

$$\begin{aligned}
\mathbb{M}\tilde{\iota} &= \mathbb{M}k_* - J'(AG)^{-1}A(g_* - \mathbb{E}^{F_*}[g_*(U)g_*(U)'](V^{-1} - V^{-1}G(G'V^{-1}G)^{-1}G'V^{-1})g_*) \\
&= \mathbb{M}k_* - J'(AG)^{-1}A(g_* - V(V^{-1} - V^{-1}G(G'V^{-1}G)^{-1}G'V^{-1})g_*) \\
&= \mathbb{M}k_* - J'(G'V^{-1}G)^{-1}G'V^{-1}g_* \\
&= \iota\,,
\end{aligned}$$

and so $\left.\frac{d\kappa(F_t)}{dt}\right|_{t=0} = \mathbb{E}^{F_*}[\iota(U)\mathbb{M}b(U)]$.

As $\phi(x) = \frac{1}{2}(x-1)^2$ for $x \geq 0$, a Taylor series expansion of $\upsilon(x)$ around $x = 0$ yields

$$D_\phi(F_t \| F_*) = \frac{t^2}{2}\mathbb{E}^{F_*}[(\mathbb{M}b(U))^2] + o(t^2)\,.$$

Therefore, whenever $\mathbb{E}^{F_*}[(\mathbb{M}b(U))^2] \neq 0$ we have

$$\frac{(\kappa(F_t) - \kappa(F_{-t}))^2}{4D_\phi(F_t\|F_*)} = \frac{\mathbb{E}^{F_*}[\iota(U)\mathbb{M}b(U)]^2 + o(1)}{\frac{1}{2}\mathbb{E}^{F_*}[(\mathbb{M}b(U))^2] + o(1)}$$

hence

$$s \geq \frac{\mathbb{E}^{F_*}[\iota(U)\mathbb{M}b(U)]^2}{\frac{1}{2}\mathbb{E}^{F_*}[(\mathbb{M}b(U))^2]}\,.$$

If $\iota(u) = 0$ (F_*-almost everywhere) then the right-hand side is zero for any b and we trivially have $s \geq 2\mathbb{E}^{F_*}[\iota(U)^2]$. Otherwise, choosing $b = \iota$ yields $s \geq 2\mathbb{E}^{F_*}[\iota(U)^2]$.

Step 2: We now prove the reverse inequality $s \leq 2\mathbb{E}^{F_*}[\iota(U)^2]$ by contradiction. Suppose there exists a sequence $\{\delta_n\}_{n\geq 1}$ with $\delta_n \downarrow 0$ and $\varepsilon > 0$ such that

$$\frac{(\overline{\kappa}_{\delta_n} - \underline{\kappa}_{\delta_n})^2}{4\delta_n} \geq 2\mathbb{E}^{F_*}[\iota(U)^2] + 2\varepsilon\,.$$

for each n. We may then choose $\underline{\theta}_n, \overline{\theta}_n \in \Theta$ and $\underline{F}_n, \overline{F}_n \in \mathcal{N}_{\delta_n}$ such that $\underline{F}_n$ and $\overline{F}_n$ satisfy $\mathbb{E}^{\overline{F}_n}[g(U, \overline{\theta}_n)] = 0$ and $\mathbb{E}^{\underline{F}_n}[k(U, \underline{\theta}_n)] = 0$, and

$$\frac{(\mathbb{E}^{\overline{F}_n}[k(U, \overline{\theta}_n)] - \mathbb{E}^{\underline{F}_n}[k(U, \underline{\theta}_n)])^2}{4\delta_n} \geq 2\mathbb{E}^{F_*}[\iota(U)^2] + \varepsilon\,. \tag{68}$$

As Θ is compact, (taking a subsequence if necessary) we can assume that $\underline{\theta}_n \to \underline{\theta}^*$ and $\overline{\theta}_n \to \overline{\theta}^*$ for some $\underline{\theta}^*, \overline{\theta}^* \in \Theta$.

The spaces $\mathcal{L}$ and $\mathcal{E}$ are equivalent to $L^2(F_*)$ for $\phi(x) = \frac{1}{2}(x-1)^2$. Let $\|\cdot\|_2$ denote the $L^2(F_*)$ norm. Note $\mathbb{E}^{F_*}[\phi(m(U))] = \frac{1}{2}\|m - 1\|_2^2$ where $m - 1$ is the function $u \mapsto m(u) - 1$. Let $\underline{m}_n$ and $\overline{m}_n$ denote the Radon–Nikodym derivatives of $\underline{F}_n$ and $\overline{F}_n$ with respect to F_*. As $\underline{F}_n, \overline{F}_n \in \mathcal{N}_{\delta_n}$, we have

$$\|\underline{m}_n - 1\|_2^2, \|\overline{m}_n - 1\|_2^2 \leq 2\delta_n \downarrow 0 \quad \text{as } n \to \infty. \tag{69}$$

By similar arguments to the proof of Lemma E.7, we may deduce $\mathbb{E}^{F_*}[g(U, \underline{\theta}^*)] = \mathbb{E}^{F_*}[g(U, \overline{\theta}^*)] = 0$. It then follows by identifiability of θ_* that $\underline{\theta}^* = \overline{\theta}^* = \theta_*$.

By differentiability of $\theta \mapsto \mathbb{E}^{F_*}[g(u, \theta)]$ at θ_*, we may deduce

$$-G(\underline{\theta}_n - \theta_*) + o(\|\underline{\theta}_n - \theta_*\|) = \mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)g(U, \underline{\theta}_n)] \quad \text{as } \underline{\theta}_n \to \theta_*.$$

It follows by Cauchy–Schwartz and the fact that G has full rank that $\|\underline{\theta}_n - \theta_*\| = O(\|\underline{m}_n - 1\|_2)$. Therefore, by (69), Cauchy–Schwarz, and $L^2(F_*)$ continuity of $\theta \mapsto g(\cdot, \theta, \gamma_0, P_{20})$ at θ_*,

$$-G(\underline{\theta}_n - \theta_*) = \mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)g_*(U)] + o(\delta_n^{1/2}) \tag{70}$$

and so

$$\underline{\theta}_n - \theta_* = -(G'V^{-1}G)^{-1}G'V^{-1}\mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)g_*(U)] + o(\delta_n^{1/2})\,.$$

Turning to the counterfactual, by similar arguments we may deduce

$$\begin{aligned}
\mathbb{E}^{F_*}[\underline{m}_n(U)k(U, \underline{\theta}_n)] - \kappa_* &= J'(\underline{\theta}_n - \theta_*) + \mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)(k_*(U) - \kappa_*)] + o(\delta_n^{1/2}) \\
&= -J'(G'V^{-1}G)^{-1}G'V^{-1}\mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)g_*(U)] \\
&\quad + \mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)(k_*(U) - \kappa_*)] + o(\delta_n^{1/2})\,.
\end{aligned}$$

However, by (70) and definition of $\mathbb{M}$ we also have

$$\mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)(k_*(U) - \kappa_* - \mathbb{M}(k_*(U) - \kappa_*))] = o(\delta_n^{1/2})$$

hence

$$\mathbb{E}^{F_*}[\underline{m}_n(U)k(U, \underline{\theta}_n)] - \kappa_* = \mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)\iota(U)] + o(\delta_n^{1/2})\,.$$

Analogous arguments apply to $\overline{m}_n$ and $\overline{\theta}_n$. We have therefore shown

$$\frac{(\mathbb{E}^{\overline{F}_n}[k(U, \overline{\theta}_n)] - \mathbb{E}^{\underline{F}_n}[k(U, \underline{\theta}_n)])^2}{4\delta_n} = \frac{(\mathbb{E}^{F_*}[(\overline{m}_n(U) - \underline{m}_n(U))\iota(U)])^2}{4\delta_n} + o(1)\,. \tag{71}$$

Note that we must have $\overline{m}_n \neq \underline{m}_n$ for all n sufficiently large. Otherwise, substituting (71) into (68) yields $o(1) \geq 2\mathbb{E}^{F_*}[\iota(U)^2] + \varepsilon$, a contradiction. Now observe that

$$\|\overline{m}_n - \underline{m}_n\|_2^2 \leq 2\|\overline{m}_n - 1\|_2^2 + 2\|\underline{m}_n - 1\|_2^2 \leq 8\delta_n \tag{72}$$

by (69). Substituting (71) and (72) into (68) yields

$$\frac{2(\mathbb{E}^{F_*}[(\overline{m}_n(U) - \underline{m}_n(U))\iota(U)])^2}{\|\overline{m}_n - \underline{m}_n\|_2^2} + o(1) \geq 2\mathbb{E}^{F_*}[\iota(U)^2] + \varepsilon\,.$$

So by Cauchy–Schwarz:

$$2\mathbb{E}^{F_*}[\iota(U)^2] + o(1) \geq 2\mathbb{E}^{F_*}[\iota(U)^2] + \varepsilon\,.$$

As $n \to \infty$, the ε term dominates the $o(1)$ term and we obtain a contradiction. ■

Proof of Lemma 7.1. Immediate by consistency of $(\hat{\theta}, \hat{\gamma}, \hat{P})$ and Slutsky's theorem. ■

E.7 Proofs for Appendix A

Proof of Proposition A.1. The minimization problem is additively separable across each $x \in \mathcal{X}$. The proof for each x follows identical arguments to the proof of Proposition 2.1. ■

Proof of Proposition A.2. This follows by straightforward modification of the proof of Proposition 2.1. ■

Proof of Proposition A.3. We prove only the result for $\underline{K}^{ex}_\delta$; the result for $\overline{K}^{ex}_\delta$ follows similarly.

Dropping dependence of k and g on (θ,γ) to simplify notation, we write (46) as

$$\begin{aligned}\underline{K}^{ex}_\delta(\theta;\gamma,P) &= \inf_{F\in\mathcal{N}^{ex}_\delta} \mathbb{E}^F[k^{ex}(U)] \quad \text{subject to} \quad \mathbb{E}^F[g^{ex}_1(U)] \le P_1\,,\ \ldots,\ \mathbb{E}^F[g^{ex}_4(U)] = 0 \\ &\ge \inf_{F\in\mathcal{N}_\delta} \mathbb{E}^F[k^{ex}(U)] \quad \text{subject to} \quad \mathbb{E}^F[g^{ex}_1(U)] \le P_1\,,\ \ldots,\ \mathbb{E}^F[g^{ex}_4(U)] = 0\,, \quad (73)\end{aligned}$$

where the first line uses Π-exchangeability and the second uses $\mathcal{N}^{ex}_\delta \subseteq \mathcal{N}_\delta$. If there exists $F \in \mathcal{N}^{ex}_\delta$ under which (1) holds at (θ,γ,P), then $\mathbb{E}^F[g^{ex}_1(U)] \le P_1$, …, $\mathbb{E}^F[g^{ex}_4(U)] = 0$ also hold under F, in which case (73) has the dual representation (48) by Proposition 2.1. Thus we have shown that the right-hand side of (48) is a lower bound for $\underline{K}^{ex}_\delta(\theta;\gamma,P)$.

We now show that a solution to (73) is Π-exchangeable. If there exists $F \in \mathcal{N}^{ex}_\delta$ under which (1) holds at (θ,γ,P), then arguing as in the proof of Lemma E.1 we know that the value of (73) is finite, say $\underline{\kappa}^\dagger$, and is attained by some $F^\dagger \in \mathcal{N}_\delta$. For each $n \ge 1$, consider

$$\inf_F D_\phi(F\|F_*) \quad \text{s.t.} \quad \mathbb{E}^F[k^{ex}(U)] \le \underline{\kappa}^\dagger + \frac{1}{n}, \mathbb{E}^F[g^{ex}_1(U)] \le P_1\,,\ \ldots,\ \mathbb{E}^F[g^{ex}_4(U)] = 0\,. \quad (74)$$

As Condition S holds for the moments g^{ex} at (θ,γ,P) and $\mathbb{E}^{F^\dagger}[k^{ex}(U)] < \underline{\kappa}^\delta + \frac{1}{n}$, it follows by similar arguments to Lemma E.4 that Condition S holds at (θ,γ,P) for the moments in (74). Therefore, by Lemma F.3 the primal problem (74) and its dual

$$\max_{\zeta\in\mathbb{R},\lambda_k\in\mathbb{R}_+,\lambda_g\in\Lambda} -\mathbb{E}^{F_*}\left[\phi^\star(-\zeta-\lambda_k k^{ex}(U)-\lambda_g' g^{ex}(U))\right]-\zeta-\lambda_{g,12}'P-\lambda_k'(\underline{\kappa}^\dagger+n^{-1})$$

are equal and the set of dual solutions is nonempty. By similar arguments to Proposition 3.2, we can deduce that the solution F_n to (74) has the form $\mathrm{d}F_n = m_n \mathrm{d}F_*$ where

$$m_n(u) = \dot\phi^\star\left(-\underline{\zeta}-\underline{\lambda}_k k^{ex}(u)-\underline{\lambda}_g' g^{ex}(u)\right)\,,$$

where $(\underline{\zeta},\underline{\lambda}_k,\underline{\lambda}_g)$ is a solution to the dual program. As Π is a group, we have $k^{ex}(M_\varpi u) = k^{ex}(u)$ and $g^{ex}(M_\varpi u) = g^{ex}(u)$ for all $\varpi \in \Pi$ and all $u \in \mathcal{U}$. As F_* is Π-exchangeable, it follows that each F_n is exchangeable.

As $D_\phi(F_n\|F_*) \le D_\phi(F^\dagger\|F_*) \le \delta$, $\{m_n\}_{n\ge1}$ is $\|\cdot\|_\phi$-norm bounded (by Lemma D.2) and therefore has a subsequence $\{m_{n_l}\}_{l\geq 1}$ converging $\mathcal{E}$-weakly to $m_0 \in \mathcal{L}$ (see Appendix D). Similar arguments to the proof of Lemma E.7 imply that the distribution F_0 given by $\mathrm{d}F_0 = m_0 \mathrm{d}F_*$ solves (73). By Π-exchangeability of each F_n, for any $\varpi \in \Pi$, measurable $A \subseteq \mathcal{U}$, and $l \geq 1$, we have $\mathbb{E}^{F_{n_l}}[\mathbb{1}\{U \in A\}] = \mathbb{E}^{F_{n_l}}[\mathbb{1}\{M_\varpi U \in A\}]$. It follows by $\mathcal{E}$-weak convergence that $\mathbb{E}^{F_0}[\mathbb{1}\{U \in A\}] = \mathbb{E}^{F_0}[\mathbb{1}\{M_\varpi U \in A\}]$, proving Π-exchangeability of F_0.

Finally, if the value of (49) is $+\infty$ then, by weak duality, so too must be the value of (73). The result follows by noting that the value of (73) provides a lower bound for (46). ∎

E.8 Proofs for Appendix B

Proof of Lemma B.1. Clearly $\underline{\kappa}_\delta \geq \inf \mathcal{K}_\infty$ for each $\delta > 0$. First suppose that $\inf \mathcal{K}_\infty$ is finite. Fix any $\varepsilon > 0$. Then there is $F_\varepsilon \in \mathcal{N}_\infty$ and $\theta_\varepsilon \in \Theta$ such that (1) holds at $(\theta_\varepsilon, \gamma_0, P_0)$ under F_ε and $\mathbb{E}^{F_\varepsilon}[k(U, \theta_\varepsilon, \gamma_0)] < \inf \mathcal{K}_\infty + \varepsilon$. Then for any $\delta \geq D_\phi(F_\varepsilon \| F_0)$ we have $\underline{\kappa}_\delta < \inf \mathcal{K}_\infty + \varepsilon$. Conversely, $\inf \mathcal{K}_\infty = -\infty$, then for each $n \in \mathbb{N}$ there exists $F_n \in \mathcal{N}_\infty$ and $\theta_n \in \Theta$ such that (1) holds at $(\theta_n, \gamma_0, P_0)$ under F_n and $\mathbb{E}^{F_n}[k(U, \theta_n, \gamma_0)] < -n$. But then for any $\delta \geq D_\phi(F_n \| F_0)$ we necessarily have $\underline{\kappa}_\delta < -n$. ∎

Proof of Lemma B.2. We prove the result only for $\underline{K}_\infty$; the result for $\overline{K}_\infty$ follows similarly.

We follow similar arguments to the proof of Proposition 2.1. Dropping dependence of on (θ, γ), consider

$$\inf_F \mathbb{E}^F[k(U)] \quad \text{subject to} \quad \mathbb{E}^F[g_1(U)] \leq P_1\,, \ldots, \mathbb{E}^F[g_4(U)] = 0\,. \tag{75}$$

Identifying distributions $F \in \mathcal{N}_\infty$ with their Radon–Nikodym derivatives $m \in \mathcal{L}$ (see Appendix D), we may define $\varphi_\infty : \mathcal{L} \times \mathbb{R}^{d+1} \to \mathbb{R} \cup \{+\infty\}$ by

$$\varphi_\infty(m, y) = \langle m, k\rangle + \mathbb{I}_{C_+}(m) + \mathbb{I}_{C_2}\left(\langle m, 1\rangle - 1 + y_1, \langle m, g\rangle - \vec{P} + y_2\right),$$

where $y_1 \in \mathbb{R}$, $y_2 \in \mathbb{R}^d$, $\mathbb{I}_C : \mathbb{R}^{d+1} \to \mathbb{R} \cup \{+\infty\}$ is given by

$$\mathbb{I}_{C_2}(y_1, y_2) = \left[\begin{array}{ll} 0 & \text{if } y_1 = 0, \text{ and } y_2 \in \mathbb{R}_-^{d_1} \times \{0\}^{d_2} \times \mathbb{R}_-^{d_3} \times \{0\}^{d_4}\,, \\ +\infty & \text{otherwise}\,, \end{array}\right.$$

and $\mathbb{I}_{C_+} : \mathcal{L} \to \mathbb{R} \cup \{+\infty\}$ is given by

$$\mathbb{I}_{C_+}(m) = \left[\begin{array}{ll} 0 & \text{if } m \geq 0 \ F_*\text{-almost everywhere}\,, \\ +\infty & \text{otherwise}\,. \end{array}\right.$$

The primal problem is $\min_{m \in \mathcal{L}} \varphi_\infty(m, y)$ and its value is $v_\infty(y) = \inf_{m \in \mathcal{L}} \varphi_\infty(m, y)$.

Under Assumption Φ, one may verify by similar arguments to Lemma E.1 that φ_∞ is proper and convex, and by similar arguments to Lemma E.2 that $v_\infty : \mathbb{R}^{d+1} \to \mathbb{R} \cup \{+\infty\}$ is proper, convex, and its effective domain is the set $\mathcal{Y}_2$ defined in display (58).

The dual problem we wish to characterize is $\max_{y^\star \in \mathbb{R}^{d+1}} (y'y^\star - \varphi_\infty^\star(0, y^\star))$ at $y = 0$, where now $\varphi_\infty^\star : \mathcal{E} \times \mathbb{R}^{d+1} \to \mathbb{R} \cup \{+\infty\}$ is the convex conjugate of φ_∞. By direct calculation,

$$\begin{aligned}
\varphi_\infty^\star(0, y^\star) &= \sup_{(m,y)\in\mathcal{L}\times\mathbb{R}^{d+1}} \Big(y'y^\star - \langle m, k\rangle - \mathbb{I}_{C_+}(m) - \mathbb{I}_{C_2}\Big(\langle m, 1\rangle - 1 + y_2, \langle m, g\rangle - \vec{P} + y_3\Big)\Big) \\
&= \sup_{m\in\mathcal{L}} \left(- y_1^\star(\langle m, 1\rangle - 1) - y_2^{\star\prime}\Big(\langle m, g\rangle - \vec{P}\Big) - \langle m, k\rangle - \mathbb{I}_{C_+}(m) \right) + \mathbb{I}_{C_2^o}(y^\star)
\end{aligned}$$

where $C_2^o = \mathbb{R}\times\Lambda$ is the polar cone of $C_2 := \operatorname{dom} \mathbb{I}_{C_2}$. Note that the supremum is never achieved at a $m \in \mathcal{L} \setminus \mathcal{L}_+$, as we can do strictly better by choosing $m = 1$ F_*-a.e. Write $y^\star \in C^o$ as $y^\star = (\zeta, \lambda)$, where $\zeta \in \mathbb{R}$ and $\lambda \in \Lambda$. By decomposibility of $\mathcal{L}$ (Rockafellar and Wets, 1998, Definition 14.59 and Theorem 14.60):

$$\begin{aligned}
\varphi_\infty^\star(0, (\zeta, \lambda)) &= \sup_{m\in\mathcal{L}_+} \mathbb{E}^{F_*}\left[m(U)(-k(U) - \zeta - \lambda' g(U))\right] + \zeta + \lambda_{12}' P \\
&= \mathbb{E}^{F_*}\left[\sup_{x\geq 0} x(-k(U) - \zeta - \lambda' g(U))\right] + \zeta + \lambda_{12}' P \\
&= \left[\begin{array}{ll} \zeta + \lambda_{12}' P & \text{if } \zeta + F_*\text{-ess}\inf(k + \lambda' g) \geq 0\,, \\ +\infty & \text{otherwise}\,, \end{array}\right.
\end{aligned}$$

provided $(\zeta, \lambda) \in C^o$. The dual value is therefore

$$\begin{aligned}
&\sup_{\zeta\in\mathbb{R},\lambda\in\Lambda} -\zeta - \lambda_{12}' P \quad \text{subject to } \zeta + F_*\text{-ess}\inf(k + \lambda' g) \geq 0 \\
&= \sup_{\lambda\in\Lambda : F_*\text{-ess}\inf(k+\lambda' g) > -\infty} \left(F_*\text{-ess}\inf(k + \lambda' g) - \lambda_{12}' P\right)\,. \qquad (76)
\end{aligned}$$

By Propositions 2.147 and 2.148(iii) of Bonnans and Shapiro (2000), we obtain a version of Lemma E.3(ii): if $0 \in \operatorname{ri}(\mathcal{Y}_2)$, then (75) and (76) are equal and the set of dual solutions is nonempty. Finally, we have that $0 \in \operatorname{ri}(\mathcal{Y}_2)$ under Condition S by Lemma E.4(ii). ■

Lemma B.3 is proved by applying results of Csiszár and Matúš (2012) that extend classical duality results relying on paired function classes to broader classes of functions. Their results apply to optimization problems constrained by equality restrictions. Straightforward modifications are required to show similar characterizations apply under inequality restrictions.

Proof of Lemma B.3. We prove the result only for $\underline{K}_{np}$; the result for $\overline{K}_{np}$ follows similarly.

We drop dependence of g and k on (θ, γ_0) in what follows. Let $\dot{\mathcal{M}} = \{m \in L^1(\mu) : \int mg \,\mathrm{d}\mu$ is finite$\}$ and $\dot{\mathcal{M}}_+ = \{m \in \dot{\mathcal{M}} : m \geq 0$ μ-a.e.$\}$. Then $F \in \mathcal{F}_\theta$ if and only if its Radon–Nikodym derivative with respect to μ, say m, belongs to $\dot{\mathcal{M}}_+$. Similarly, any $m \in \mathcal{M}_+$ with $\int m \,\mathrm{d}\mu = 1$

corresponds to a distribution in $\mathcal{F}_\theta$. Let $\vec{P}_0 = (P_0, 0_{d_3+d_4})$. For any $y = (y_1, y_2) \in \mathbb{R} \times \mathbb{R}^d$, let

$$\mathcal{M}[y] = \left\{ m \in \dot{\mathcal{M}}_+ : \int m \, \mathrm{d}\mu = 1 + y_1 \, , \int m g \, \mathrm{d}\mu = \vec{P}_0 + y_2 \right\} ,$$

where the second integration is element-wise. Define $q_1, q_2 : \mathbb{R}^{d+1} \to \mathbb{R} \cup \{+\infty\}$ by

$$q_1(y) = \inf_{m \in \mathcal{M}[y]} \int m k \, \mathrm{d}\mu$$

with the understanding that $q_1(y) = +\infty$ if the infimum runs over an empty set, and

$$q_2(y) = \left[\begin{array}{cl} 0 & \text{if } y \in \{0\} \times \mathbb{R}_-^{d_1} \times \{0\}^{d_2} \times \mathbb{R}_-^{d_3} \times \{0\}^{d_4} \, , \\ +\infty & \text{otherwise.} \end{array} \right.$$

The function q_2 is proper and convex. Moreover, q_1 is convex and proper because μ-essential boundedness of k guarantees that $q_1(y) > -\infty$ for all y.

Note that $\operatorname{dom} q_1 = \{(\int m \, \mathrm{d}\mu - 1, (\int m g \, \mathrm{d}\mu - \vec{P}_0)) : m \in \dot{\mathcal{M}}_+\}$. Under condition S_{np}, by similar arguments to the proof of Lemma E.4(ii) and Corollary 6.6.2 of Rockafellar (1970), we have

$$\begin{aligned} (1, \vec{P}_0) &\in \operatorname{ri}(\{\textstyle\int m(1, g) \, \mathrm{d}\mu : m \in \dot{\mathcal{M}}_+\} + \{0\} \times \mathcal{C}) \\ &= \operatorname{ri}(\{(\textstyle\int m(1, g) \, \mathrm{d}\mu : m \in \dot{\mathcal{M}}_+\}) + \operatorname{ri}(\{0\} \times \mathcal{C}) \, . \end{aligned}$$

Equivalently, $0 \in \operatorname{ri}(\operatorname{dom} q_1) + \operatorname{ri}(\{0\} \times \mathcal{C})$ and so $\operatorname{ri}(\operatorname{dom} q_1) \cap \operatorname{ri}(\operatorname{dom} q_2)$ is nonempty. Then by Fenchel's Duality Theorem (Rockafellar, 1970, Theorem 31.1),

$$\underline{K}_{np}(\theta; \gamma_0, P_0) = \inf_{y \in \mathbb{R}^{d+1}} (q_1(y) + q_2(y)) = \sup_{y^\star \in \mathbb{R}^{d+1}} (-q_1^\star(y^\star) - q_2^\star(-y^\star)) \, , \tag{77}$$

where $q_1^\star$ and $q_2^\star$ are the convex conjugates of q_1 and q_2. Write $y^\star = (\zeta, \lambda)$. Then

$$q_2^\star(-(\zeta, \lambda)) = \left[\begin{array}{cl} 0 & \text{if } -\lambda \in \Lambda \, , \\ +\infty & \text{otherwise.} \end{array} \right. \tag{78}$$

For $q_1^\star$, we begin by writing

$$\begin{aligned}
-q_1^\star((\zeta,\lambda)) &= \inf_{y\in\mathbb{R}^{d+1}} \big(-(\zeta,\lambda')y + q_1(y)\big) \\
&= \inf_{y\in\mathbb{R}^{d+1}} \inf_{m\in\mathcal{M}[y]} \left(-(\zeta,\lambda')y + \int mk \,\mathrm{d}\mu\right) \\
&= \inf_{y\in\mathbb{R}^{d+1}} \inf_{m\in\mathcal{M}[y]} \left(\zeta + \lambda'\vec{P}_0 + \int \big(k-\zeta-\lambda' g\big)\, m \,\mathrm{d}\mu\right) \\
&= \inf_{m\in\dot{\mathcal{M}}_+} \left(\zeta + \lambda'\vec{P}_0 + \int \big(k-\zeta-\lambda' g\big)\, m \,\mathrm{d}\mu\right) .
\end{aligned}$$

Let

$$Q(u,m(u)) = \left[\begin{array}{ll} k(u)m(u) & \text{if } m(u)\geq 0\,, \\ +\infty & \text{otherwise.} \end{array}\right.$$

We therefore have

$$-q_1^\star((\zeta,\lambda)) = \inf_{m\in\dot{\mathcal{M}}} \left(\zeta + \lambda'\vec{P}_0 + \int \big(Q(u,m(u)) - \zeta - \lambda' g(u)\big)\, m(u) \,\mathrm{d}\mu(u)\right) .$$

By Remark A.3 and Theorem A.4 of Csiszár and Matúš (2012), we may bring the inf inside the integral:

$$\begin{aligned}
-q_1^\star((\zeta,\lambda)) &= \zeta + \lambda'\vec{P}_0 + \int \inf_{x\geq 0} \big(k(u)-\zeta-\lambda' g(u)\big)\, x \,\mathrm{d}\mu(u) \\
&= \left[\begin{array}{cl} -\infty & \text{if } \mu\text{-ess}\inf(k-\zeta-\lambda' g) < 0\,, \\ \zeta + \lambda'\vec{P}_0 & \text{otherwise.} \end{array}\right.
\end{aligned} \tag{79}$$

Letting λ_{12} denote the first $d_1 + d_2$ elements of λ, it now follows from (77), (78), and (79) that

$$\begin{aligned}
\underline{K}_{np}(\theta;\gamma_0,P_0) &= \sup_{\zeta\in\mathbb{R},\lambda\in\Lambda:\mu\text{-ess}\inf(k-\zeta+\lambda' g)\geq 0} \zeta - \lambda_{12}' P_0 \\
&= \sup_{\lambda\in\Lambda:\mu\text{-ess}\inf(k+g'\lambda) > -\infty} \mu\text{-ess}\inf(k+\lambda' g) - \lambda_{12}' P_0
\end{aligned}$$

as required. ■

E.9 Proofs for Appendix D

Proof of Lemma D.2. Follows by taking $c = \frac{1}{2}$ in the definition of $\|\cdot\|_\phi$. ■

Proof of Lemma D.3. By Lemma D.2, it suffices to show $\mathbb{E}^{F_*}[\phi(m(U))] < \infty$ for all $m \in \mathcal{L}_+$. As ϕ satisfies the Δ_2-condition under Assumption Φ(i), $m \in \mathcal{L}$ implies $\mathbb{E}^{F_*}[\phi(1+c|m(U)|)] < \infty$ for all $c > 0$. As F_* is a finite measure, $\mathcal{L}$ also contains constant functions. As $\mathcal{L}$ is closed under addition, we therefore have

$$\infty > \mathbb{E}^{F_*}[\phi(1+|m(U)-1|)] = \mathbb{E}^{F_*}[\phi(m(U))1\!\!1\{m(U) \geq 1\}] + \mathbb{E}^{F_*}[\phi(2-m(U))1\!\!1\{m(U) \leq 1\}]$$

which implies $\mathbb{E}^{F_*}[\phi(m(U))1\!\!1\{m(U) \geq 1\}]$ is finite. Finiteness of $\mathbb{E}^{F_*}[\phi(m(U))1\!\!1\{m(U) \leq 1\}]$ follows because $\infty > \phi(0) \geq \phi(x) \geq \phi(1) = 0$ for $x \in [0,1]$ under Assumption Φ(i). ■

F Supplementary Results

F.1 Notation

For $x \in \mathbb{R}^n$ and $A \subset \mathbb{R}^n$ let $d(x,A) = \inf_{a \in A} \|x-a\|$. Let $\vec{d}_H(A,B) = \sup_{a \in A} \inf_{b \in B} \|a-b\|$ denote the directed Hausdorff distance between sets $A, B \subset \mathbb{R}^n$. Let B_ε denote a Euclidean ball centered at the origin with radius ε, where the dimension of the ball should be obvious from the context. Let $T \subseteq \mathbb{R}^n$ be a nonempty, closed convex cone with nonempty interior. Let $\partial A = \mathrm{cl}(A) \setminus \mathrm{int}(A)$ denote the boundary of $A \subset T$ (relative to $\mathbb{R}^n$) and $\partial(A;T) = \mathrm{cl}(\partial A \cap \mathrm{int}(T))$ denote is boundary relative to T. For example, $T = \mathbb{R}_+ \times \mathbb{R}$, and $A = \{(x,y) \in T : x^2+y^2 \leq 1\}$, then $\partial A = \{(x,y) \in T : x^2+y^2 = 1\} \cup \{0\} \times [-1,1]$ and $\partial(A;T) = \{(x,y) \in T : x^2+y^2 = 1\}$.

F.2 Stability of Constraint Qualifications under Perturbations

Lemma F.1 *Let Assumption Φ hold and let Condition S' hold at (θ,γ,P). Then: there exists a neighborhood N of P such that Condition S' holds at $(\theta,\gamma,\tilde{P})$ for each $\tilde{P} \in N$.*

Proof of Lemma F.1. By condition S', there exists $\varepsilon > 0$ such that

$$B_{2\varepsilon} \subseteq \left(\{\mathbb{E}^F[g(U,\theta,\gamma)] - (P, 0_{d_3+d_4}) : F \in \mathcal{N}_\infty\} + \mathcal{C}\right).$$

Then for any $\tilde{P}$ with $\|P - \tilde{P}\| < \varepsilon$, we have

$$\|(\mathbb{E}^F[g(U,\theta,\gamma)] - (P, 0_{d_3+d_4})) - (\mathbb{E}^F[g(U,\theta,\gamma)] - (\tilde{P}, 0_{d_3+d_4}))\| < \varepsilon$$

for all $F \in \mathcal{N}_\infty$, and so $B_\varepsilon \subseteq (\{\mathbb{E}^F[g(U,\theta,\gamma)] - (\tilde{P}, 0_{d_3+d_4}) : F \in \mathcal{N}_\infty\} + \mathcal{C})$. ■

Lemma F.2 *Let Assumption Φ hold, let each entry of g be $\mathcal{E}$-continuous in (θ,γ), and let Condition S' hold at (θ,γ,P). Then: there exists a neighborhood N of (θ,γ,P) such that Condition S' holds at $(\tilde{\theta},\tilde{\gamma},\tilde{P})$ for each $(\tilde{\theta},\tilde{\gamma},\tilde{P}) \in N$.*

Proof of Lemma F.2. As Condition S' holds at (θ,γ,P), there exists sufficiently large δ such that $0 \in \mathrm{int}(\{\mathbb{E}^F[g(U,\theta,\gamma)] - (P, 0_{d_3+d_4}) : F \in \mathcal{N}_\delta\} + \mathcal{C})$. (To see this, fit a sufficiently small hypercube around 0, identify a density $F \in \mathcal{N}_\infty$ with each vertex, and take δ to be the largest ϕ-divergence from F_* of each of the densities at the vertex.) Therefore, we may choose $\varepsilon > 0$ such that $B_{4\varepsilon} \subseteq \mathrm{int}(\{\mathbb{E}^F[g(U,\theta,\gamma)] - (P, 0_{d_3+d_4}) : F \in \mathcal{N}_\delta\} + \mathcal{C})$.

Identify any $F \in \mathcal{N}_\delta$ with its Radon–Nikodym derivative with respect to F_*, say m. The set of all such m, say $\mathcal{M}_\delta$, is a $\|\cdot\|_\phi$-bounded subset of $\mathcal{L}$ (Lemma D.2). By $\mathcal{E}$-continuity, there exists a neighborhood N_1 of (θ,γ) such that for any $(\tilde\theta,\tilde\gamma) \in N_1$ and with r denoting any entry of $g_1,\ldots,g_4$, we have

$$\|r(\cdot,\theta,\gamma) - r(\cdot,\tilde\theta,\tilde\gamma)\|_\psi < \frac{\varepsilon}{\sqrt{d}(2+\phi(2)+\delta)}\,.$$

It follows by Hölder's inequality for Orlicz classes and Lemma D.2 that

$$\sup_{m\in\mathcal{M}_\delta} |\mathbb{E}^{F_*}[m(U)r(U,\theta,\gamma)] - \mathbb{E}^{F_*}[m(U)r(U,\tilde\theta,\tilde\gamma)]| \le \frac{\varepsilon}{\sqrt{d}}$$

for any $(\tilde\theta,\tilde\gamma) \in N_1$. Let N_2 be an ε-neighborhood of P. For any $F \in \mathcal{N}_\delta$ and any $(\tilde\theta,\tilde\gamma,\tilde P) \in N_1 \times N_2$, we have

$$\|(\mathbb{E}^F[m(U)g(U,\theta,\gamma)] - (P,0)) - (\mathbb{E}^F[m(U)g(U,\tilde\theta,\tilde\gamma)] - (\tilde P,0))\| < 2\varepsilon\,,$$

hence

$$\begin{aligned} B_{2\varepsilon} &\subseteq \mathrm{int}(\{\mathbb{E}^F[g(U,\tilde\theta,\tilde\gamma)] - (\tilde P, 0_{d_3+d_4}) : F \in \mathcal{N}_\delta\} + \mathcal{C}) \\ &\subseteq \mathrm{int}(\{\mathbb{E}^F[g(U,\tilde\theta,\tilde\gamma)] - (\tilde P, 0_{d_3+d_4}) : F \in \mathcal{N}_\infty\} + \mathcal{C})\,, \end{aligned}$$

as required. ■

F.3 Additional Details on the Program $\Delta^\star(\theta;\gamma,P)$

We derive the dual of (17) using similar arguments to Appendix E.2. Suppressing dependence of k and g on (θ,γ) to simplify notation, define $\varphi : \mathcal{L}\times\mathbb{R}^{d+1} \to \mathbb{R}\cup\{+\infty\}$ by

$$\varphi(m,y) = Q_\phi(m) + \mathbb{I}_{C_2}\left(\langle m,1\rangle - 1 + y_1, \langle m,g\rangle - \vec{P} + y_2\right)\,,$$

where $y = (y_1,y_2) \in \mathbb{R}\times\mathbb{R}^d$ and $\mathbb{I}_{C_2} : \mathbb{R}^{d+1} \to \mathbb{R}\cup\{+\infty\}$ is given by

$$\mathbb{I}_{C_2}(y_1,y_2) = \left[\begin{array}{ll} 0 & \text{if } y_1 = 0, \text{ and } y_2 \in \mathbb{R}^{d_1}_- \times \{0\}^{d_2} \times \mathbb{R}^{d_3}_- \times \{0\}^{d_4}\,, \\ +\infty & \text{otherwise}\,. \end{array}\right.$$

For any $y \in \mathbb{R}^{d+1}$, define the primal problem

$$\min_{m\in\mathcal{L}} \varphi(m,y) \qquad (\mathrm{P}'_y)$$

and let $v(y) = \inf_{m \in \mathcal{L}} \varphi(m, y)$ denote its value. Then $v(0)$ is the value in the original problem (17). By similar arguments to Lemmas E.1 and E.2 we may deduce that φ and v are proper and convex and that the effective domain of v is the set $\mathcal{Y}_2$ defined in (58). Existence of a solution to the primal problem (P'_y) for each $y \in \mathcal{Y}_2$ follows similarly, and uniqueness of the solution follows by strict convexity of ϕ.

By similar arguments to Appendix E.2, the dual problem of (P'_y) is

$$\max_{\zeta \in \mathbb{R}, \lambda \in \Lambda} \zeta y_1 + \lambda' y_2 - \mathbb{E}^{F_*}\Big[\phi^\star(-\zeta - \lambda' g(U))\Big] - \zeta - \lambda'_{12} P \,. \tag{D'_y}$$

The dual of (17), which corresponds to the dual (D'_y) with $y = 0$, is $\Delta^\star(\theta; \gamma, P)$ from (16).

Lemma F.3 *Let Assumption Φ hold. Then: the dual program of $\Delta(\theta; \gamma, P)$ is $\Delta^\star(\theta; \gamma, P)$ in (16). If Condition S holds at (θ, γ, P) then the primal and dual values are equal and the set of solutions to the dual problem is nonempty and convex. Moreover, if Condition S' holds, then the set of dual solutions is compact.*

Proof of Lemma F.3. The dual program is derived above, which requires only Assumption Φ. Equality of the primal and dual programs and existence of dual solutions when $0 \in \mathrm{ri}(\mathcal{Y}_2)$ follows by similar arguments to Lemma E.3(ii). Lemma E.4 shows that Conditions S and S' are sufficient for $0 \in \mathrm{ri}(\mathcal{Y}_2)$ and $0 \in \mathrm{int}(\mathcal{Y}_2)$, respectively. Compactness of the set of dual solutions when $0 \in \mathrm{int}(\mathcal{Y}_2)$ follows by similar arguments to Lemma E.3(iii). ■

Lemma F.4 *Let Assumption Φ hold, let $\mathbb{E}^{F_*}[\phi^\star(c_1 + c_2' g(U, \theta, \gamma))]$ be continuous in (θ, γ) for every $(c_1, c_2) \in \mathbb{R} \times \mathbb{R}^d$. Then: the function $(\theta, \gamma, P) \mapsto \Delta^\star(\theta; \gamma, P)$ is continuous at any point (θ, γ, P) at which Condition S' holds.*

Proof of Lemma F.4. Fix some (θ, γ, P) at which Conditions S' holds. Note we must have $\Delta^\star(\theta; \gamma, P) < \infty$. The objective function

$$(\zeta, \lambda) \mapsto L(\zeta, \lambda; \theta, \gamma, P) := -\mathbb{E}^{F_*}[\phi^\star(-\zeta - \lambda' g(U, \theta, \gamma))] - \zeta - \lambda'_{12} P$$

is the pointwise infimum of affine functions of (ζ, λ) and is therefore concave and u.s.c. By Lemma F.3, it has a nonempty, convex, and compact set of maximizers $\Xi \subset \mathbb{R} \times \Lambda$. Fix $\varepsilon > 0$ and let $\Xi^\varepsilon = \{(\zeta, \lambda) \in \mathbb{R} \times \Lambda : d((\zeta, \lambda), \Xi) \leq \varepsilon\}$.

By continuity of $(\theta, \gamma) \mapsto \mathbb{E}^{F_*}[\phi^\star(c_1 + c_2' g(U, \theta, \gamma))]$, for any $(\zeta, \lambda) \in \mathbb{R}^{d+1}$ we have

$$L(\zeta, \lambda; \tilde{\theta}, \tilde{\gamma}, \tilde{P}) \to L(\zeta, \lambda; \theta, \gamma, P) \quad \text{as } (\tilde{\theta}, \tilde{\gamma}, \tilde{P}) \to (\theta, \gamma, P) \,.$$

By concavity of L in its first two arguments, convergence may be strengthened to hold uniformly over the compact set Ξ^ε (Rockafellar, 1970, Theorem 10.8) and so, in particular,

$$\sup_{(\zeta,\lambda)\in\Xi^\varepsilon} L(\zeta,\lambda;\tilde\theta,\tilde\gamma,\tilde P) \to \Delta^\star(\theta;\gamma,P) \quad \text{as } (\tilde\theta,\tilde\gamma,\tilde P) \to (\theta,\gamma,P)\,. \tag{80}$$

By u.s.c. of $L(\cdot;\theta,\gamma,P)$ and definition of Ξ:

$$\Delta^\star(\theta;\gamma,P) > \sup_{(\zeta,\lambda)\in\partial(\Xi^\varepsilon;\mathbb{R}\times\Lambda)} L(\zeta,\lambda;\theta,\gamma,P)\,. \tag{81}$$

By (80) and (81) there exists a neighborhood N of (θ,γ,P) such that for any $(\tilde\theta,\tilde\gamma,\tilde P)\in N$,

$$\sup_{(\zeta,\lambda)\in\Xi^\varepsilon} L(\zeta,\lambda;\tilde\theta,\tilde\gamma,\tilde P) > \sup_{(\zeta,\lambda)\in\partial(\Xi^\varepsilon;\mathbb{R}\times\Lambda)} L(\zeta,\lambda;\tilde\theta,\tilde\gamma,\tilde P)$$

holds. By standard arguments for maximizers of convex objective functions (e.g. Theorem 2.7 of Newey and McFadden (1994)), whenever $(\tilde\theta,\tilde\gamma,\tilde P)\in N$ we have that

$$\Delta^\star(\tilde\theta;\tilde\gamma,\tilde P) := \sup_{(\zeta,\lambda)\in\mathbb{R}\times\Lambda} L(\zeta,\lambda;\tilde\theta,\tilde\gamma,\tilde P) = \sup_{(\zeta,\lambda)\in\Xi^\varepsilon} L(\zeta,\lambda;\tilde\theta,\tilde\gamma,\tilde P)\,.$$

The result now follows by (80). ■

F.4 Additional Details on Convergence of Multipliers

Recall $\underline{\Xi}_\delta(\theta;\gamma,P)$ and $\overline{\Xi}_\delta(\theta;\gamma,P)$ denote the sets of multipliers (η,ζ,λ) solving (13) and (14). By Lemma E.5, these sets are nonempty, convex, and compact whenever Condition S' holds at (θ,γ,P) and there exists $F\in\mathcal{N}_\delta$ with $D_\phi(F\|F_*)<\delta$ such that (1) holds under F at (θ,γ,P). In particular, this is true if Condition S' holds at (θ,γ,P) and $\Delta^\star(\theta;\gamma,P)<\delta$.

Let $T=\mathbb{R}_+\times\mathbb{R}\times\Lambda$. Provided $\underline{\Xi}_\delta(\theta;\gamma,P)$ and $\overline{\Xi}_\delta(\theta;\gamma,P)$ are compact, for each $\varepsilon>0$ we may cover $\underline{\Xi}_\delta(\theta;\gamma,P)$ and $\underline{\Xi}_\delta(\theta;\gamma,P)$ by sets $\underline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon\subset T$ and $\overline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon\subset T$ consisting of finitely many hypercubes with edges parallel to the coordinate axes so that

$$d((\eta,\zeta,\lambda),\underline{\Xi}_\delta(\theta;\gamma,P)) \le \varepsilon \text{ for all } (\eta,\zeta,\lambda)\in\underline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon$$
$$d((\eta,\zeta,\lambda),\overline{\Xi}_\delta(\theta;\gamma,P)) \le \varepsilon \text{ for all } (\eta,\zeta,\lambda)\in\overline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon$$

and so that $\partial(\underline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon;T)\cap\underline{\Xi}_\delta(\theta;\gamma,P)=\emptyset$ and $\partial(\overline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon;T)\cap\overline{\Xi}_\delta(\theta;\gamma,P)=\emptyset$.

Lemma F.5 *Let Assumptions Φ and M(i)(ii) hold, let Condition S hold at (θ,γ,P), and let $\Delta^\star(\theta;\gamma,P)<\delta$. Then:*

(i) $\underline{K}^\star_\delta$ and $\overline{K}^\star_\delta$ are continuous at (θ,γ,P);

(ii) for each $\varepsilon>0$ there exists a neighborhood N of (θ,γ,P) such that $\underline{\Xi}_\delta(\tilde\theta;\tilde\gamma,\tilde P)\subseteq\underline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon$ and $\overline{\Xi}_\delta(\tilde\theta;\tilde\gamma,\tilde P)\subseteq\overline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon$ for each $(\tilde\theta,\tilde\gamma,\tilde P)\in N$;

(iii) $\vec{d}_H(\underline{\Xi}_\delta(\tilde{\theta};\tilde{\gamma},\tilde{P}),\underline{\Xi}_\delta(\theta;\gamma,P)) \to 0$ *and* $\vec{d}_H(\overline{\Xi}_\delta(\tilde{\theta};\tilde{\gamma},\tilde{P}),\overline{\Xi}_\delta(\theta;\gamma,P)) \to 0$ *as* $(\tilde{\theta},\tilde{\gamma},\tilde{P}) \to (\theta,\gamma,P)$.

Proof of Lemma F.5. We prove the result for $\underline{K}^\star_\delta$ and $\underline{\Xi}_\delta$; the result for $\overline{K}^\star_\delta$ and $\overline{\Xi}_\delta$ follows similarly.

Step 1: Preliminaries. To simplify notation, let $\Xi = \underline{\Xi}_\delta(\theta;\gamma,P)$ and $\Xi^\varepsilon = \underline{\Xi}_\delta(\theta;\gamma,P)^\varepsilon$. Lemmas F.2 and F.4 imply there is a neighborhood N' of (θ,γ,P) such that Condition S' holds at $(\tilde{\theta},\tilde{\gamma},\tilde{P})$ and $\Delta^\star(\tilde{\theta},\tilde{\gamma},\tilde{P}) < \delta$ for each $(\tilde{\theta},\tilde{\gamma},\tilde{P}) \in N'$. By Lemma E.5, the multipliers $\underline{\Xi}_\delta(\tilde{\theta};\tilde{\gamma},\tilde{P})$ solving the program $\underline{K}^\star_\delta(\tilde{\theta};\tilde{\gamma},\tilde{P})$ are a nonempty, convex, compact subset of T for each $(\tilde{\theta},\tilde{\gamma},\tilde{P}) \in N'$.

The objective function

$$(\eta,\zeta,\lambda) \mapsto L(\eta,\zeta,\lambda;\theta,\gamma,P) := -\mathbb{E}^{F_*}[(\eta\phi)^\star(-k(U,\theta,\gamma) - \zeta - \lambda' g(U,\theta,\gamma))] - \eta\delta - \zeta - \lambda_{12}' P$$

is the pointwise infimum of affine functions of (η,ζ,λ) and is therefore concave and u.s.c. By u.s.c. and definition of Ξ, we have

$$\underline{K}^\star_\delta(\theta;\gamma,P) - \sup_{(\eta,\zeta,\lambda)\in\partial(\Xi^\varepsilon;T)} L(\eta,\zeta,\lambda;\theta,\gamma,P) =: 4a > 0\,. \tag{82}$$

The remaining steps of the proof depend on whether or not $\min\{\eta : (\eta,\zeta,\lambda) \in \Xi\} > 0$.

Step 2: Proof of parts (i) and (ii) when $\min\{\eta : (\eta,\zeta,\lambda) \in \Xi\} > 0$. W.l.o.g. we may choose Ξ^ε such that $\min\{\eta : (\eta,\zeta,\lambda) \in \Xi^\varepsilon\} > 0$. For any $\eta > 0$,

$$L(\eta,\zeta,\lambda;\tilde{\theta},\tilde{\gamma},\tilde{P}) = -\eta\mathbb{E}^{F_*}\left[\phi^\star\left(\frac{k(U,\tilde{\theta},\tilde{\gamma})+\zeta+\lambda' g(U,\tilde{\theta},\tilde{\gamma})}{-\eta}\right)\right] - \eta\delta - \zeta - \lambda_{12}'\tilde{P}\,.$$

By Assumption M(ii), for any $(\eta,\zeta,\lambda) \in (0,\infty)\times\mathbb{R}^{d+1}$ we have

$$L(\eta,\zeta,\lambda;\tilde{\theta},\tilde{\gamma},\tilde{P}) \to L(\eta,\zeta,\lambda;\theta,\gamma,P) \quad \text{as } (\tilde{\theta},\tilde{\gamma},\tilde{P}) \to (\theta,\gamma,P)\,.$$

By concavity of L in (η,ζ,λ), convergence may be strengthened to hold uniformly over the compact set Ξ^ε (Rockafellar, 1970, Theorem 10.8) and so, in particular,

$$\sup_{(\eta,\zeta,\lambda)\in\Xi^\varepsilon} L(\eta,\zeta,\lambda;\tilde{\theta},\tilde{\gamma},\tilde{P}) \to \underline{K}^\star_\delta(\theta;\gamma,P) \text{ as } (\tilde{\theta},\tilde{\gamma},\tilde{P}) \to (\theta,\gamma,P). \tag{83}$$

By (82) and (83), there exists a neighborhood N'' of (θ,γ,P) such that for $(\tilde{\theta},\tilde{\gamma},\tilde{P}) \in N'\cap N''$, the inequality

$$\sup_{(\eta,\zeta,\lambda)\in\Xi^\varepsilon} L(\eta,\zeta,\lambda;\tilde{\theta},\tilde{\gamma},\tilde{P}) > \sup_{(\eta,\zeta,\lambda)\in\partial(\Xi^\varepsilon;T)} L(\eta,\zeta,\lambda;\tilde{\theta},\tilde{\gamma},\tilde{P})$$

holds. By similar arguments to the proof of Lemma F.4 we may deduce that $\Xi(\tilde{\theta};\tilde{\gamma},\tilde{P}) \subseteq \Xi^{\varepsilon}$ holds on $N := N' \cap N''$, proving part (ii). Continuity (part (i)) now follows by (83).

Step 3: Proof of part (ii) when $\min\{\eta : (\eta,\zeta,\lambda) \in \Xi\} = 0$. Take $(\bar{\zeta},\bar{\lambda}) \in \mathbb{R} \times \Lambda$ such that $(0,\bar{\zeta},\bar{\lambda}) \in \Xi$. As in the proof of Lemma E.5, we may deduce

$$\lim_{\eta \downarrow 0} L(\eta,\bar{\zeta},\bar{\lambda};\theta,\gamma,P) = L(0,\bar{\zeta},\bar{\lambda};\theta,\gamma,P) = K_{\delta}^{\star}(\theta;\gamma,P)\,.$$

For any $\varepsilon_0 \in (0,a)$, we may choose $\bar{\eta} > 0$ such that $L(\bar{\eta},\bar{\zeta},\bar{\lambda};\theta,\gamma,P) > K_{\delta}^{\star}(\theta;\gamma,P) - \varepsilon_0$ and $(\bar{\eta},\bar{\zeta},\bar{\lambda}) \in \mathrm{int}(\Xi^{\varepsilon})$. By Assumption M(ii), there exists a neighborhood N'' of (θ,γ,P) upon which

$$L(\bar{\eta},\bar{\zeta},\bar{\lambda};\tilde{\theta},\tilde{\gamma},\tilde{P}) > \underline{K}_{\delta}^{\star}(\theta;\gamma,P) - 2\varepsilon_0 \tag{84}$$

holds for all $(\tilde{\theta},\tilde{\gamma},\tilde{P}) \in N''$. We now argue by contradiction that the inequality

$$\sup_{(\eta,\zeta,\lambda)\in\partial(\Xi^{\varepsilon};T)} L(\eta,\zeta,\lambda;\theta,\gamma,P) \geq \sup_{(\eta,\zeta,\lambda)\in\partial(\Xi^{\varepsilon};T)} L(\eta,\zeta,\lambda;\tilde{\theta},\tilde{\gamma},\tilde{P}) - 2\varepsilon_0 \tag{85}$$

holds on a neighborhood, say N''', of (θ,γ,P).

Suppose that there is $\varepsilon_1 > 0$ and $(\theta_n,\gamma_n,P_n) \to (\theta,\gamma,P)$ along which

$$\sup_{(\eta,\zeta,\lambda)\in\partial(\Xi^{\varepsilon};T)} L(\eta,\zeta,\lambda;\theta,\gamma,P) \leq \sup_{(\eta,\zeta,\lambda)\in\partial(\Xi^{\varepsilon};T)} L(\eta,\zeta,\lambda;\theta_n,\gamma_n,P_n) - \varepsilon_1\,. \tag{86}$$

For each $n \geq 1$, choose $(\eta_n,\zeta_n,\lambda_n) \in \arg\sup_{(\eta,\zeta,\lambda)\in\partial(\Xi^{\varepsilon};T)} L(\eta,\zeta,\lambda;\theta_n,\gamma_n,P_n)$. As $\partial(\Xi^{\varepsilon};T)$ is compact, take a subsequence $\{(\eta_{n_l},\zeta_{n_l},\lambda_{n_l})\}_{l\geq 1}$ converging to $(\eta^*,\zeta^*,\lambda^*) \in \partial(\Xi^{\varepsilon};T)$. If $\eta^* > 0$, then by uniform convergence of $L(\cdot;\theta_n,\gamma_n,P_n)$ to $L(\cdot;\theta,\gamma,P)$ on compact subsets of $(0,\infty) \times \mathbb{R} \times \mathbb{R}^d$, we obtain

$$\lim_{l\to\infty} L(\eta_{n_l},\zeta_{n_l},\lambda_{n_l};\theta_{n_l},\gamma_{n_l},P_{n_l}) \leq \sup_{(\eta,\zeta,\lambda)\in\partial(\Xi^{\varepsilon};T)} L(\eta,\zeta,\lambda;\theta,\gamma,P)\,,$$

contradicting (86). Conversely, if $\eta^* = 0$, fix any small $\varepsilon_2 > 0$ so that $(\varepsilon_2,\zeta^*,\lambda^*) \in \partial(\Xi^{\varepsilon};T)$. By u.s.c. and concavity of $L(\cdot,\zeta^*,\lambda^*;\theta,\gamma,P)$, we may choose ε_2 sufficiently small that

$$L(\varepsilon_2,\zeta^*,\lambda^*;\theta,\gamma,P) - L(2\varepsilon_2,\zeta^*,\lambda^*;\theta,\gamma,P) < \varepsilon_1\,. \tag{87}$$

For all l large enough we have $\eta_{n_l} < \varepsilon_2$ and hence $\tau_{n_l} := \frac{\varepsilon_2}{2\varepsilon_2 - \eta_{n_l}} \in (0,1)$. By concavity,

$$L(\varepsilon_2,\zeta_{n_l},\lambda_{n_l};\theta_{n_l},\gamma_{n_l},P_{n_l}) \geq \tau_{n_l} L(\eta_{n_l},\zeta_{n_l},\lambda_{n_l};\theta_{n_l},\gamma_{n_l},P_{n_l}) + (1-\tau_{n_l}) L(2\varepsilon_2,\zeta_{n_l},\lambda_{n_l};\theta_{n_l},\gamma_{n_l},P_{n_l})\,.$$

which rearranges to yield

$$L(\eta_{n_l}, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \leq \frac{1}{\tau_{n_l}} \left(L(\varepsilon_2, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) - (1-\tau_{n_l}) L(2\varepsilon_2, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \right) .$$

By uniform convergence of $L(\cdot; \theta_n, \gamma_n, P_n)$ to $L(\cdot; \theta, \gamma, P)$ on compact subsets of $(0,\infty) \times \mathbb{R}^{d+1}$, we obtain

$$\lim_{l \to \infty} L(\eta_{n_l}, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \leq 2L(\varepsilon_2, \zeta^*, \lambda^*; \theta, \gamma, P) - L(2\varepsilon_2, \zeta^*, \lambda^*; \theta, \gamma, P) .$$

It follows by (87) that for all l sufficiently large we have

$$L(\eta_{n_l}, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) < \sup_{(\eta,\zeta,\lambda) \in \partial(\Xi^\varepsilon; T)} L(\eta, \zeta, \lambda; \theta, \gamma, P) + \varepsilon_1 ,$$

which contradicts (86). This completes the proof of inequality (85).

It now follows from displays (82), (84), and (85) that on $N' \cap N'' \cap N'''$ we have

$$\begin{aligned} L(\bar\eta, \bar\zeta, \bar\lambda; \tilde\theta, \tilde\gamma, \tilde P) > \underline{K}^\star_\delta(\theta; \gamma, P) - 2\varepsilon_0 &= \sup_{(\eta,\zeta,\lambda) \in \partial(\Xi^\varepsilon; T)} L(\eta, \zeta, \lambda; \theta, \gamma, P) + 4a - 2\varepsilon_0 \\ &\geq \sup_{(\eta,\zeta,\lambda) \in \partial(\Xi^\varepsilon; T)} L(\eta, \zeta, \lambda; \tilde\theta, \tilde\gamma, \tilde P) + 4(a - \varepsilon_0) . \end{aligned}$$

As $a - \varepsilon_0 > 0$, we have $\sup_{(\eta,\zeta,\lambda) \in \Xi^\varepsilon} L(\eta, \zeta, \lambda; \tilde\theta, \tilde\gamma, \tilde P) > \sup_{(\eta,\zeta,\lambda) \in \partial(\Xi^\varepsilon; T)} L(\eta, \zeta, \lambda; \tilde\theta, \tilde\gamma, \tilde P)$ holds on $N := N' \cap N'' \cap N'''$. By similar arguments to the proof of Lemma F.4 we may deduce that $\Xi(\tilde\theta; \tilde\gamma, \tilde P) \subseteq \Xi^\varepsilon$ holds on N, proving part (ii).

Step 4: Proof of part (i) when $\min\{\eta : (\eta, \zeta, \lambda) \in \Xi\} = 0$. For any $(\tilde\theta, \tilde\gamma, \tilde P) \in N$, we have

$$\underline{K}^\star_\delta(\tilde\theta; \tilde\gamma, \tilde P) = \sup_{(\eta,\zeta,\lambda) \in \Xi^\varepsilon} L(\eta, \zeta, \lambda; \tilde\theta, \tilde\gamma, \tilde P) \tag{88}$$

by Step 3, so

$$\underline{K}^\star_\delta(\tilde\theta; \tilde\gamma, \tilde P) \geq L(\bar\eta, \bar\zeta, \bar\lambda; \tilde\theta, \tilde\gamma, \tilde P) > \underline{K}^\star_\delta(\theta; \gamma, P) - 2\varepsilon_0 ,$$

by (84), proving l.s.c. To establish u.s.c., for each $\varepsilon_0 > 0$ one may deduce (by similar arguments used to establish (85) in Step 3) there is a neighborhood N'''' of (θ, γ, P) upon which

$$\sup_{(\eta,\zeta,\lambda) \in \Xi^\varepsilon} L(\eta, \zeta, \lambda; \theta, \gamma, P) \geq \sup_{(\eta,\zeta,\lambda) \in \Xi^\varepsilon} L(\eta, \zeta, \lambda; \tilde\theta, \tilde\gamma, \tilde P) - \varepsilon_0 \tag{89}$$

holds. It follows by (88) and (89) that on $N \cap N''''$, we have

$$\underline{K}^\star_\delta(\tilde\theta; \tilde\gamma, \tilde P) = \sup_{(\eta,\zeta,\lambda) \in \Xi^\epsilon} L(\eta, \zeta, \lambda; \tilde\theta, \tilde\gamma, \tilde P) \leq \sup_{(\eta,\zeta,\lambda) \in \Xi^\epsilon} L(\eta, \zeta, \lambda; \theta, \gamma, P) + \varepsilon_0 = \underline{K}^\star_\delta(\theta; \gamma, P) + \varepsilon_0 ,$$

proving u.s.c.

Step 5: Proof of part (iii). By part (ii), for each $\varepsilon > 0$ there exists a neighborhood N such that $\underline{\Xi}_\delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P}) \subseteq \Xi_\delta(\theta; \gamma, P)^\varepsilon$ holds for all $(\tilde{\theta}, \tilde{\gamma}, \tilde{P}) \in N$. Therefore, the inequality

$$\vec{d}_H(\underline{\Xi}_\delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P}), \underline{\Xi}_\delta(\theta; \gamma, P)) \leq \vec{d}_H(\Xi_\delta(\theta; \gamma, P)^\varepsilon, \underline{\Xi}_\delta(\theta; \gamma, P)) \leq \varepsilon$$

holds for all $(\tilde{\theta}, \tilde{\gamma}, \tilde{P}) \in N$. ■

References

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer-Verlag, New York.

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

Csiszár, I. and F. Matúš (2012). Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. *Kybernetika 48*(4), 637–689.

Fang, Z. and A. Santos (2019). Inference on directionally differentiable functions. *Review of Economic Studies 86*(1), 377–412.

Hiriart-Urruty, J.-B. and C. Lemaréchal (2001). *Fundamentals of Convex Analysis.* Springer.

Komunjer, I. and G. Ragusa (2016). Existence and characterization of conditional density projections. *Econometric Theory 32*(4), 947–987.

Krasnosel'skii, M. A. and Ya. B. Rutickii (1961). *Convex Functions and Orlicz Spaces.* P. Noordhoff Ltd., Groningen.

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. Volume 4 of *Handbook of Econometrics*, pp. 2111–2245. Elsevier.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press.

Rockafellar, R. T. and R. J. B. Wets (1998). *Variational Analysis.* Springer.

Sason, I. and S. Verdú (2016). f-divergence inequalities. *IEEE Transactions on Information Theory 62*(11), 5973–6006.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research 30*(1), 169–186.