# Counterfactual Sensitivity and Robustness*

Timothy Christensen[†] Benjamin Connault[‡]

May 16, 2022

**Abstract**

We propose a framework for analyzing the sensitivity of counterfactuals to parametric assumptions about the distribution of latent variables in structural models. In particular, we derive bounds on counterfactuals as the distribution of latent variables spans nonparametric neighborhoods of a given parametric specification while other "structural" features of the model are maintained. Our approach recasts the infinite-dimensional problem of optimizing the counterfactual with respect to the distribution of latent variables (subject to model constraints) as a finite-dimensional convex program. We develop an MPEC version of our method to further simplify computation in models with endogenous parameters (e.g., value functions) defined by equilibrium constraints. We propose plug-in estimators of the bounds and two methods for inference. We also show that our bounds converge to the sharp nonparametric bounds on counterfactuals as the neighborhood size becomes large. To illustrate the broad applicability of our procedure, we present empirical applications to welfare analysis in matching models with transferable utility and dynamic discrete choice models.

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

**JEL codes:** C14, C18, C54, D81

*We are very grateful to a co-editor and four anonymous referees, R. Allen, S. Bonhomme, G. Chamberlain, A. Galvao, B. Kaplowitz, 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

[‡]Latour Trading LLC. benjamin.connault@gmail.com

# 1 Introduction

Researchers frequently make parametric assumptions about the distribution of latent variables in structural models. These assumptions are typically made for computational convenience[1] or because simulation-based methods are used for estimation. In many models, such as those we consider in this paper, the distribution of latent variables is not nonparametrically identified. This raises the possibility that model parameters and the outcomes of policy experiments, or *counterfactuals*, may be only partially identified when parametric assumptions are relaxed. That is, different distributions may fit the data equally well in-sample, but may yield different values of the counterfactual. It is therefore natural to question whether counterfactuals are sensitive or robust to researchers' parametric assumptions, especially when evaluating the credibility of structural modeling exercises.

This paper proposes a framework for analyzing the sensitivity of counterfactuals to parametric assumptions about the distribution of latent variables in structural models defined by moment (in)equalities. In particular, we derive bounds on counterfactuals as the distribution of latent variables spans nonparametric neighborhoods of the researcher's parametric specification while other "structural" features of the model are maintained. This approach is in the spirit of global sensitivity analysis advocated by Leamer (1985) (see also Tamer (2015)). Global sensitivity analyses are important in this context: many structural models are nonlinear so policy interventions can have different effects at different points in the parameter space. A major difficulty with implementing global sensitivity analyses is tractability. A more tractable alternative are local sensitivity analyses, which are based on small perturbations around a chosen specification. Because local approaches rely on linearization, they may fail to correctly characterize the range of counterfactuals predicted by a nonlinear model when the distribution differs nontrivially from the researcher's specification.

Our main insight is to borrow from the robustness literature in economics pioneered by Hansen and Sargent (2001, 2008) to simplify computation using convex programming.[2] Following this literature, we define neighborhoods around the researcher's parametric specification using statistical divergence (e.g., Kullback–Leibler divergence), with the option to add certain shape restrictions as appropriate. For tractability, we restrict our attention to models

[1] Examples include the conventional Gumbel (or type-I extreme value) assumption in discrete choice models following McFadden (1974), dynamic discrete choice models following Rust (1987), and matching models with transferable utility following Dagsvik (2000) and Choo and Siow (2006). Models of static or dynamic discrete games often impose parametric assumptions about the distribution of payoff shocks—see, e.g., Berry (1992), Aguirregabiria and Mira (2007), Bajari, Benkard, and Levin (2007), and Ciliberto and Tamer (2009).

[2] Our approach is also related to the field of *distributionally robust optimization* in operations research. See, e.g., Shapiro (2017), Duchi and Namkoong (2021), and references therein.

that may be written as a finite number of moment (in)equalities, where the expectation is with respect to the distribution of latent variables. While restrictive, this class accommodates many important models of static and dynamic discrete choice, discrete games, and matching.

To describe our procedure, consider the problem of minimizing or maximizing the counterfactual at a fixed value of structural parameters by varying the distribution of latent variables over a neighborhood, subject to the model's (in)equality restrictions. We use duality to recast this infinite-dimensional optimization problem as a finite-dimensional convex program. The value of this *inner* program is treated as a criterion function, which is optimized in an *outer* optimization with respect to structural parameters. The dimension of the inner problem is independent of the neighborhood size, making our procedure tractable over both small and large neighborhoods. To further simplify computation, we develop an MPEC version of our procedure for models featuring endogenous parameters defined by equilibrium constraints (e.g., value functions defined by Bellman equations). We show that this implementation can produce significant computational gains for dynamic discrete choice models in particular.

Our approach is conceptually different from nonparametric partial identification analyses which derive bounds on counterfactuals under minimal distributional assumptions. But as we show, bounds computed using our procedure converge to the (sharp) nonparametric bounds in the limit as the neighborhood size becomes large. Aside from sensitivity analyses, our methods may therefore be used to approximate nonparametric bounds by taking the neighborhood size to be large but finite.

For estimation and inference, we propose simple plug-in estimators of the bounds and establish their consistency. We also propose and theoretically justify two methods for inference: a computationally simple but conservative projection procedure and a more efficient bootstrap procedure.

We illustrate our procedure with two empirical applications. The first revisits the "marital college premium" estimates reported in Chiappori, Salanié, and Weiss (2017), which relied on an i.i.d. Gumbel (type-I extreme value) assumption for the distribution of individuals' idiosyncratic marital preferences (see also Choo and Siow (2006)). The second empirical application performs a counterfactual welfare analysis in the canonical dynamic discrete choice model of Rust (1987).

**Related literature.** Our approach has connections with global prior sensitivity in Bayesian analysis (Chamberlain and Leamer, 1976; Leamer, 1982; Berger, 1984), most notably Giacomini, Kitagawa, and Uhlig (2016) and Ho (2018) who 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 semiparametric likelihood models using sieve approximations for the infinite-dimensional nuisance parameter (the distribution of latent variables in our setting). For the class of moment-based models we consider, our approach instead eliminates the infinite-dimensional nuisance parameter via a convex program of fixed dimension.

Several other works have used convex duality to characterize identified sets in models with latent variables. Most closely related are Ekeland, Galichon, and Henry (2010) and Schennach (2014).[3] The problem we study is different, both because of its focus on counterfactuals, rather than structural parameters, and because the optimization is performed over a neighborhood, rather than over all distributions. As a consequence, our estimation and inference methods are also quite different.

Torgovitsky (2019b) uses linear programming to characterize sharp identified sets in latent variable models defined by quantile restrictions. Within this class, his approach is more computationally convenient than ours for characterizing identified sets. Several important moments or counterfactuals cannot be expressed as quantile restrictions, such as social surplus in discrete choice models and Bellman equations in dynamic discrete choice models. Our approach is compatible with these, and therefore allows for characterizing identified sets in broader classes of model, as well as performing sensitivity analyses.

There is also a literature deriving nonparametric bounds on counterfactuals in specific latent variables models. Examples include Manski (2007, 2014), Allen and Rehbeck (2019), Tebaldi, Torgovitsky, and Yang (2019), Lafférs (2019), Torgovitsky (2019a), and Gualdani and Sinha (2020). Most closely related is Norets and Tang (2014), who construct identified sets of counterfactual conditional choice probabilities (CCPs) in dynamic binary choice models. Their approach is specific to counterfactual CCPs and to dynamic binary choice models. We allow for a wider range of counterfactuals (e.g., welfare), shape restrictions, and multinomial choice, as well as performing sensitivity analyses.[4]

Finally, our work is complementary to the recent literature on local sensitivity—see, e.g., Kitamura, Otsu, and Evdokimov (2013), Andrews, Gentzkow, and Shapiro (2017, 2020), Armstrong and Kolesár (2021), Bonhomme and Weidner (2021), and Mukhin (2018). Much of this literature is concerned with local misspecification of moment conditions, which is different from the setting we consider.

[3]Works using other notions of "duality" to construct identified sets include Beresteanu, Molchanov, and Molinari (2011), Galichon and Henry (2011), Chesher and Rosen (2017), and Li (2018).

[4]Kalouptsidi, Scott, and Souza-Rodrigues (2021) 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.

**Outline.** Section 2 introduces our procedure, estimators of the bounds, and shows our approach recovers nonparametric bounds as the neighborhood size becomes large. Section 3 discusses practical aspects and implementation details. Section 4 gives guidance for interpreting the neighborhood size. Empirical applications are presented in Section 5. Section 6 discusses estimation and inference. The online appendix presents extensions of our methodology, connections with local sensitivity analyses, additional empirical results, and proofs of our main results. A secondary online appendix presents background material on Orlicz classes and supplemental proofs.

# 2 Procedure

We begin in Section 2.1 by describing the class of models to which our procedure may be applied. Section 2.2 describes our approach, Section 2.3 shows how duality is used to simplify the bounds, and Section 2.4 introduces our estimators of the bounds. Section 2.5 shows our bounds converge to the sharp nonparametric bounds as the neighborhood size becomes large.

## 2.1 Setup

We consider a class of models that link 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

$$\mathbb{E}^F[g_1(U,\theta,\gamma_0)] \leq P_{10}, \tag{1a}$$
$$\mathbb{E}^F[g_2(U,\theta,\gamma_0)] = P_{20}, \tag{1b}$$
$$\mathbb{E}^F[g_3(U,\theta,\gamma_0)] \leq 0, \tag{1c}$$
$$\mathbb{E}^F[g_4(U,\theta,\gamma_0)] = 0, \tag{1d}$$

where $g_1, \ldots, g_4$ are vectors of moment functions, $P_0 = (P_{10}, P_{20})$ is partitioned conformably, and $\mathbb{E}^F$ denotes expectation with respect to a vector of latent variables $U \sim F$. We assume that the researcher has consistent estimators $(\hat{P}, \hat{\gamma})$ of $(P_0, \gamma_0)$. We also assume that the researcher is interested in a (scalar) counterfactual of the form

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

This setup accommodates counterfactuals that do not depend explicitly on $U$, in which case (2) reduces to $\kappa = k(\theta, \gamma_0)$. Note that $\kappa$ will still depend on the distribution of $U$ through $\theta$, whose values are disciplined by the moment conditions (1).

Several models and counterfactuals of interest fall into this framework. We review three examples before proceeding.

**Example 2.1 (Discrete choice and consumer welfare)** Suppose an individual derives utility $h_j(X, \theta) + U_j$ from choice $j \in \mathcal{J}_0 := \{0, 1, \ldots, J\}$, where $X \in \mathcal{X}$ are observed covariates and $U = (U_j)_{j \in \mathcal{J}_0}$ is latent (to the econometrician). We assume, as typical, that $U$ is drawn independently across individuals from a continuous distribution $F$. The probability that an individual with characteristics $x$ chooses $j$ is

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

where $\mathbb{P}_F$ denotes probabilities when $U \sim F$. In empirical work, $\theta$ is typically estimated using a criterion that fits the model-implied choice probabilities (3) to probabilities observed in the data. 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 is the average utility consumers with characteristics $x$ derive from the choice problem. Another common welfare measure is the change in surplus $\Delta W(x_a, x_b) = W(x_a) - W(x_b)$ associated with a shift from $x_b$ to $x_a$. In practice, it is common to assume the $U_j$ are i.i.d. Gumbel (type-I extreme value), as this yields closed-form expressions for choice probabilities, $W(x)$, and $\Delta W(x_a, x_b)$.

Our approach may be used to perform a sensitivity analysis of $W(x)$ and $\Delta W(x_a, x_b)$ to parametric assumptions about $F$ when $\mathcal{X}$ is finite. A leading example is matching models with finitely many agent types—see Section 5.1 and references therein. Understanding the sensitivity of $W(x)$ and $\Delta W(x_a, x_b)$ to $F$ is important in this case because $W(x)$ and $\Delta W(x_a, x_b)$ are not nonparametrically identified.[5]

In our notation, $g_2$ collects indicator functions representing the choice probabilities (3) across covariates $x \in \mathcal{X}$ and choices $j \in \mathcal{J} := \{1, \ldots, J\}$ ($j = 0$ is redundant):

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

[5] See, e.g., Berry and Haile (2010, 2014) and Allen and Rehbeck (2019) for nonparametric identification of utilities and welfare measures in discrete choice models when characteristics have continuous support.

and $P_{20} = (\Pr(j|x))_{(j,x)\in\mathcal{J}\times\mathcal{X}}$ is the vector of true choice probabilities. There are no $g_1$, $g_3$, $g_4$, or $\gamma$ in this model. Finally, $k(U,\theta) = \max_{j\in\mathcal{J}_0}(h_j(x,\theta)+U_j)$ for $W(x)$ and $k(U,\theta) = \max_{j\in\mathcal{J}_0}(h_j(x_a,\theta)+U_j) - \max_{j\in\mathcal{J}_0}(h_j(x_b,\theta)+U_j)$ for $\Delta W(x_a,x_b)$. □

**Example 2.2 (Discrete games)** Following Bresnahan and Reiss (1990, 1991), Berry (1992), and Tamer (2003), consider the complete-information game in Table 1.

| | | Firm 2: 0 | Firm 2: 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 $X = x$.

Here $U = (U_1, U_2)$ is the latent (to the econometrician) component of firms' profits, which is 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 $X$) from data on a large number of markets. As the model is incomplete—there are values of $U$ for which there are multiple equilibria—moment inequality methods are typically used in empirical work to avoid restricting the equilibrium selection mechanism. However, strong parametric assumptions are often made about the distribution of $U$ (typically bivariate Normal) to derive the model-implied probabilities for different market structures; see, e.g., Berry (1992), Ciliberto and Tamer (2009), Beresteanu et al. (2011), and Kline and Tamer (2016). It therefore 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 $(1,0)$ across $x \in \mathcal{X}$, with $P_{10}$ denoting the corresponding true probabilities. The inequalities are typically expressed as upper bounds on the probabilities of $(0,1)$ and $(1,0)$; we flip the sign to be compatible with (1a):

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

where $\theta = (\Delta_1, \Delta_2, \beta_1, \beta_2)$. Similarly, $g_2$ and $P_{20}$ collect the moment conditions and probabil-

[6]Continuous regressors are often discretized in empirical applications; see, e.g., Ciliberto and Tamer (2009), Grieco (2014), Kline and Tamer (2016), and Chen, Christensen, and Tamer (2018).

ities for outcomes $(0,0)$ and $(1,1)$, which are always realized as the result of unique equilibria:

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

There is no $g_3$, $g_4$, or $\gamma$ in this model. Ciliberto and Tamer (2009) compute upper bounds on the probability of entrants under a counterfactual payoff shift, say $\tau(\theta)$. The function $k(U,\theta) = \mathbb{1}\{U_1 \geq \tau(\theta) - \beta_1' x\}$ corresponds to the upper bound on the probability of firm 1 entering when $X = x$ under this counterfactual. □

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

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

where $s \in \mathcal{S}$ is a Markov state variable, $\mathcal{D}_0 = \{0,1,\ldots,D\}$ is the set of actions, $\pi_{d,s}$ is the flow payoff for action $d$ in state $s$ which is parameterized by $\theta_\pi$, $U_d$ is a latent utility shock, $\beta \in (0,1)$ is a discount parameter, and $E[\,\cdot\,|d,s]$ denotes expectation with respect to the future state $s'$. The distribution $F$ of $U = (U_d)_{d\in\mathcal{D}_0}$ is typically assumed to be continuous and independent of $s$. The CCP of action $d$ in state $s$ is

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

where $\mathbb{P}_F$ denotes probabilities when $U \sim F$.

It is common practice to assume the $U_d$ are i.i.d. Gumbel, as this yields closed-form expressions for the expectation in (4) and multinomial-logit expressions for the CCPs. Parameters $\theta_\pi$ or $(\theta_\pi, \beta)$ are typically estimated using a criterion function that fits the model-implied CCPs (5) to probabilities observed in the data. Counterfactuals are then computed by solving (4) under alternative laws of motion, flow payoffs, or other interventions.

When $\mathcal{S}$ is finite, model parameters, counterfactual CCPs, and counterfactual welfare measures are typically not identified without parametric restrictions on $F$. Our procedure may be used perform a sensitivity analysis of counterfactuals to parametric assumptions on $F$ as follows. Let $\theta = (\theta_\pi, v, \tilde{v})$ or $\theta = (\theta_\pi, \beta, v, \tilde{v})$, where $v = (V(s))_{s\in\mathcal{S}}$ and $\tilde{v} = (\tilde{V}(s))_{s\in\mathcal{S}}$ collect the baseline and counterfactual value functions across $s \in \mathcal{S}$. Also let $\gamma = (M_d)_{d\in\mathcal{D}_0}$ collect the transition matrices for $s$, $g_2$ collect indicator functions for the CCPs (5) across states $s \in \mathcal{S}$

and choices $d \in \mathcal{D} := \{1, \ldots, D\}$ ($d = 0$ is redundant):

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

with $M_{d,s}$ denoting the $s$th row of $M_d$, and $P_{20} = (\Pr(d|s))_{(d,s)\in\mathcal{D}\times\mathcal{S}}$ collect the corresponding true CCPs. Finally, $g_4$ collects moment functions representing (4) in the baseline model and under the counterfactual:

$$g_4(U, \theta, \gamma) = \begin{bmatrix} (\max_{d \in \mathcal{D}_0}\{\pi_{d,s}(\theta_\pi) + U_d + \beta M_{d,s} v\} - v_s)_{s \in \mathcal{S}} \\ (\max_{d \in \mathcal{D}_0}\{\tilde{\pi}_{d,s}(\theta_\pi) + U_d + \tilde{\beta} \tilde{M}_{d,s} \tilde{v}\} - \tilde{v}_s)_{s \in \mathcal{S}} \end{bmatrix}, \quad (6)$$

where $v_s = V(s)$, $\tilde{v}_s = \tilde{V}(s)$, and $\tilde{\pi}$, $\tilde{\beta}$, $\tilde{M}_d$ denote counterfactual flow payoffs, discount factor, and law of motion.[7] We recommend including the location normalizations $\mathbb{E}^F[U_d] = 0$ for $d \in \mathcal{D}_0$ in $g_4$ for interpretability. We also recommend including scale normalizations in $g_4$ so that $\mathbb{E}^F[\max_{d\in\mathcal{D}_0} U_d]$ is finite. For instance, in Section 5.2 we normalize $\mathbb{E}^F[U_d^2]$ for all $d \in \mathcal{D}_0$.

Counterfactual CCPs can be computed using

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

Change in average welfare corresponds to $k(\theta, \gamma) = w'(\tilde{v} - v)$ for a weight vector $w$. □

**Remark 2.1** *We allow for conditional moments models with* $\mathbb{E}[g_1(U, X, \theta, \gamma)|X = x] \leq P_{10}(x)$ *(and similarly for (1b)-(1d)) if $U$ is independent of $X$ and $X$ takes values in a finite set $\mathcal{X}$. Moment functions are then stacked across $x \in \mathcal{X}$ to form $g_1$, $g_2$, $g_3$, and $g_4$ (see Examples 2.1-2.3). Appendix A discusses extensions to conditional moment models where the distribution of $U$ may vary with the value of (discrete) covariates, and to non-separable models with discrete covariates. Models with continuous covariates fall outside the scope of our procedure.*

[7] If $\mathbb{E}^F[\max_{d\in\mathcal{D}_0} U_d]$ is finite, then $v \mapsto (\mathbb{E}^F[\max_{d\in\mathcal{D}_0}\{\pi_{d,s}(\theta_\pi) + U_d + \beta M_{d,s} v\}])_{s\in\mathcal{S}}$ is a $\ell^\infty$-contraction of modulus $\beta$ on $\mathbb{R}^{|\mathcal{S}|}$. Hence, there is a unique $(v, \tilde{v})$ solving $\mathbb{E}^F[g_4(U, \theta, \gamma)] = 0$ at any fixed $(\theta_\pi, \beta, \tilde{\beta}, F)$. The solution $(v, \tilde{v})$ must collect the solutions to (4) in the baseline model and counterfactual across states: $v = (V(s))_{s\in\mathcal{S}}$ and $\tilde{v} = (\tilde{V}(s))_{s\in\mathcal{S}}$. It follows that $F$ satisfies $\mathbb{E}^F[g_4(U, \theta, \gamma)] = 0$ at $\theta = (\theta_\pi, \beta, v, \tilde{v})$ if and only if $(v, \tilde{v})$ corresponds to the value functions $V$ and $\tilde{V}$ under $F$.

**Remark 2.2** *Our setup relies on the counterfactual being expressible as (2). If $k$ is vector-valued, our procedure can be applied to compute the support function*[8] *of the identified set of counterfactuals: set $k^\tau(U,\theta,\gamma) = \tau' k(U,\theta,\gamma)$ for a conformable unit vector $\tau$ and replace (2) with $\kappa^\tau = \mathbb{E}^F[k^\tau(U,\theta,\gamma_0)]$. Our setup excludes counterfactuals that are infinite-dimensional, such as the distribution of the number of firms in a market.*

**Remark 2.3** *The distribution $F$ is not nonparametrically identified in any of the above examples or, more generally, in the class of models (1) when the support of $U$ contains many more points than there are moment conditions (e.g., when $U$ is continuously distributed).*

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 $\hat{\gamma}$ of $\gamma_0$, the researcher computes an estimate $\hat{\theta}$ of $\theta$ using a criterion function based on the moment conditions

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

Finally, the researcher estimates the counterfactual using $\hat{\kappa} = \mathbb{E}^{F_*}[k(U,\hat{\theta},\hat{\gamma})]$. If $k$ does not depend on $U$, then the estimated counterfactual is simply $\hat{\kappa} = k(\hat{\theta},\hat{\gamma})$. In this case $\hat{\kappa}$ will still depend implicitly on $F_*$ through $\hat{\theta}$.[9]

The researcher's chosen specification $F_*$ is used both for estimation of $\theta$ and again when computing the counterfactual. A natural question is: to what extent does the counterfactual depend on the choice of distribution? The main contribution of this paper is to provide a tractable econometric framework for answering this question.

## 2.2 Our Approach

As a sensitivity analysis, we shall relax the researcher's parametric specification and allow $F$ to vary over nonparametric neighborhoods $\mathcal{N}_\delta$ of $F_*$, where $\delta$ is a measure of neighborhood "size". When we do so, there may be multiple pairs $(\theta, F) \in \Theta \times \mathcal{N}_\delta$ that satisfy (1) but which yield different values of the counterfactual. Our objects of interest are the smallest and largest

[8] A closed convex set is determined by its support function—see Rockafellar (1970, Section 13).

[9] While this discussion has assumed point identification of $\theta$ and $\kappa$ for sake of exposition, our methods allow structural parameters and counterfactuals to be partially identified.

values of the counterfactual over all such $(\theta, F)$ 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 naturally accommodates models with partially-identified structural parameters and counterfactuals. Our approach also sidesteps having to compute the identified set of structural parameters.

The optimization problems (8) and (9) are made tractable by a convenient choice of $\mathcal{N}_\delta$. Following Hansen and Sargent (2001) and Maccheroni, Marinacci, and Rustichini (2006), we consider neighborhoods constrained by $\phi$-divergence (Csiszár, 1975):

$$\mathcal{N}_\delta = \{F \in \mathcal{F} : D_\phi(F \| F_*) \leq \delta\},$$
$$D_\phi(F \| F_*) = \begin{cases} \displaystyle\int \phi\left(\frac{\mathrm{d}F}{\mathrm{d}F_*}\right) \mathrm{d}F_* & \text{if } F \ll F_*, \\ +\infty & \text{otherwise,} \end{cases} \tag{10}$$

where $\mathcal{F}$ denotes all probability measures on the support[10] $\mathcal{U}$ of $U$ and $F \ll F_*$ denotes absolute continuity of $F$ with respect to $F_*$. The convex function $\phi : [0, \infty) \to \mathbb{R}_+ \cup \{+\infty\}$ 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 $\chi^2$ divergence, and

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

corresponds to $L^p$ divergence. If $F_*$ has positive (Lebesgue) density, then the absolute continuity condition merely rules out $F$ with mass points.

**Remark 2.4** *Normalizations and other shape restrictions may be added by augmenting the moment functions $g_1, \ldots, g_4$. Examples include: (i) location normalizations, e.g. $\mathbb{E}^F[U] = 0$ or $\mathbb{E}^F[\mathbb{1}\{U_i \leq 0\} - 0.5] = 0$ for each element $U_i$ of $U$; (ii) scale normalizations, e.g. $\mathbb{E}^F[U_i^2] = 1$; (iii) covariance normalizations, e.g. $\mathbb{E}^F[UU'] = I$; and (iv) smoothness restrictions, e.g. $\mathbb{E}^F[\mathbb{1}\{U_i \leq a_{k+1}\} - \mathbb{1}\{U_i \leq a_k\}] \leq C$ for $a_1 < \ldots < a_K$ and a positive constant $C$.*

**Remark 2.5** *Appendix A.1 shows that shape restrictions including symmetry, exchangeability, and, more generally, invariance under a finite group of transforms, are also easy to impose.*

[10] That is, $\mathcal{U}$ is the set of all values that $U$ could conceivably take according to the model, which is possibly larger that the support of the measure $F_*$.

## 2.3 Dual Formulation

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 $(\gamma, P)$. To do so, we say that the moment conditions (1) hold "at $(\theta,\gamma,P)$" if they hold when $\gamma_0$ is replaced by $\gamma$ 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 $(\theta,\gamma,P)$.

We first impose some mild regularity conditions on $F_*$, $\phi$, and the moment functions to justify the dual formulation. Similar conditions are used in generalized empirical likelihood estimation (see, e.g., Komunjer and Ragusa (2016)). Let $\Phi_0$ denote the set of all $\phi : [0,\infty) \to \mathbb{R} \cup \{+\infty\}$ such that $\phi$ is continuously differentiable on $(0,+\infty)$ and strictly convex, with $\phi(1) = \phi'(1) = 0$, $\phi(0) < +\infty$, $\lim_{x\downarrow 0} \phi'(x) < 0$, $\lim_{x\to+\infty} \phi(x)/x = +\infty$, $\lim_{x\to+\infty} \phi'(x) > 0$, and $\lim_{x\to+\infty} x\phi'(x)/\phi(x) < +\infty$. The functions inducing KL, $\chi^2$, and $L^p$ divergence all belong to $\Phi_0$.

Let $\phi^\star(x) = \sup_{t\geq 0:\phi(t)<+\infty}(tx - \phi(t))$ denote the convex conjugate of $\phi \in \Phi_0$ and let $\psi(x) = \phi^\star(x) - x$. Define $\mathcal{E} = \{f : \mathcal{U} \to \mathbb{R}$ for which $\mathbb{E}^{F_*}[\psi(c|f(U)|)] < \infty$ for all $c > 0\}$, an Orlicz class of functions (see Appendix F for details). For example,

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

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** $\Phi$ *(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* $\gamma \in \Gamma$.

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 faster than exponentially (e.g., Gaussian but not Gumbel). Assumption $\Phi$ therefore fails for KL divergence in Examples 2.1 and 2.3, but holds for $\chi^2$ or $L^p$ divergence as these only require finite second or $q$th moments, respectively.

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 $\lambda_{12}$ denote the first $d_1 + d_2$ elements of $\lambda$. A derivation of the following criterion functions is presented in Appendix G.2.

**Proposition 2.1** *Suppose that Assumption* $\Phi$ *holds. Then the criterion functions (11) and (12) may be restated as*

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

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

*Moreover, the value of (13) is* $+\infty$ *(equivalently, the value of (14) is* $-\infty$*) if and only if there is no distribution in* $\mathcal{N}_\delta$ *under which (1) holds at* $(\theta, \gamma, P)$.

**Remark 2.6** *Problems (13) and (14) are convex in* $(\eta, \zeta, \lambda)$. *The parameter* $\eta$ *is the Lagrange multiplier for the constraint* $D_\phi(F\|F_*) \leq \delta$. *Similarly,* $\lambda$ *collects the Lagrange multipliers for the moment (in)equalities (1a)–(1d). These multipliers are non-negative if they correspond to inequality restrictions and unconstrained otherwise. Finally,* $\zeta$ *is the Lagrange multiplier for the constraint* $\int \mathrm{d}F = 1$, *which ensures that the optimization is over probability measures.*

Problems (13) and (14) simplify in some special cases. For KL neighborhoods, $\phi^\star(x) = e^x - 1$ and the multiplier $\zeta$ has a closed-form solution, 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 $k(u, \theta, \gamma)$ does not depend on $u$. To analyze this case, consider

$$\Delta(\theta;\gamma,P) := \inf_F D_\phi(F\|F_*) \quad \text{subject to (1) holding at } (\theta,\gamma,P). \tag{15}$$

The value $\Delta(\theta;\gamma,P)$ is the minimum $\phi$-divergence between $F_*$ and a distribution $F$ for which the moment conditions hold at $(\theta,\gamma,P)$. We show in Proposition G.2 that $\Delta(\theta;\gamma,P)$ has an equivalent dual formulation:

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

For KL divergence, $\zeta$ may be solved for in closed-form and the dual program simplifies to

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

If $k$ does not depend on $u$, then by a change of variables[11] we may restate problems (13) and (14) as

$$\underline{K}_\delta(\theta;\gamma,P) = \left[\begin{array}{l} k(\theta,\gamma) \\ +\infty \end{array}\right. ,\quad \overline{K}_\delta(\theta;\gamma,P) = \left[\begin{array}{ll} k(\theta,\gamma) & \text{if } \Delta(\theta;\gamma,P)\leq\delta, \\ -\infty & \text{if } \Delta(\theta;\gamma,P)>\delta. \end{array}\right. \tag{17}$$

An important feature of our approach is that the optimization problems (13), (14), and (16) are convex and their dimension does not increase with $\delta$. This feature is not shared by other seemingly natural approaches to flexibly model $F$, such as mixtures or other finite-dimensional sieves. As we show in Section 2.5, our procedure may be used to approximate sharp nonparametric bounds on counterfactuals by taking $\delta$ to be large but finite.

## 2.4 Estimation

We now propose simple estimators of the bounds $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ based on "plugging in" consistent estimators $(\hat{P},\hat{\gamma})$ of $(P_0,\gamma_0)$. Estimators $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are computed by optimizing criterion functions with respect to $\theta$:

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

where

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

[11] Substitute $\eta\zeta - k(\theta,\gamma)$ in place of $\zeta$ in (13) and $\eta\zeta + k(\theta,\gamma)$ in place of $\zeta$ in (14), then substitute $\eta\lambda$ in place of $\lambda$ in both (13) and (14).

and $\underline{K}_\delta(\theta;\hat{\gamma},\hat{P})$, $\overline{K}_\delta(\theta;\hat{\gamma},\hat{P})$, and $\Delta(\theta;\hat{\gamma},\hat{P})$ are the criterion functions (13), (14), and (16) evaluated at $(\hat{\gamma},\hat{P})$. If $k(u,\theta,\gamma)=k(\theta,\gamma)$, then we simply have

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

In Section 6.1 we establish consistency of $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ and derive their asymptotic distribution.

### 2.5 Nonparametric Bounds on Counterfactuals

We define the (nonparametric) identified set of counterfactuals as

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

where $\mathcal{F}_\theta=\{F\in\mathcal{F}:\mathbb{E}^F[g(U,\theta,\gamma_0)] \text{ is finite and } F\ll\mu\}$ denotes all distributions on $\mathcal{U}$ that are absolutely continuous with respect to a $\sigma$-finite dominating measure $\mu$ and for which the moments in (1) are finite at $\theta$. We impose existence of a density with respect to $\mu$ as it is often a structural assumption used, e.g., to avoid ties in CCPs or to establish existence of equilibria. The main result of this section shows that $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ approach the sharp nonparametric bounds $\inf\mathcal{K}$ and $\sup\mathcal{K}$ as $\delta$ becomes large.

We first introduce some additional regularity conditions. Say $k$ is "$\mu$-essentially bounded" if $|k(\cdot,\theta,\gamma_0)|$ has finite $\mu$-essential supremum[12] for each $\theta\in\Theta$. This holds trivially if $k$ is bounded (e.g., counterfactual CCPs in Examples 2.2 and 2.3 and change in average welfare in Example 2.3). Models with unbounded $k$ may be reparameterized (as a proof device) by setting $\tilde{\theta}=(\theta,\kappa)$, appending $k(U,\theta,\gamma_0)-\kappa$ as an element of $g_4$, and setting $k(U,\tilde{\theta},\gamma_0)=\kappa$.

We also require a constraint qualification condition. This is a sufficient condition for establishing equivalence of "nonparametric" primal and dual problems in Appendix B, which is an intermediate step in the proof of the following result. Let $0_{d_i}$ denote a $d_i\times 1$ vector of zeros, $\mathcal{C}=\mathbb{R}_+^{d_1}\times\{0_{d_2}\}\times\mathbb{R}_+^{d_3}\times\{0_{d_4}\}$, $\mathcal{G}(\theta,\gamma)=\{\mathbb{E}^F[g(U,\theta,\gamma)] : F\in\mathcal{N}_\infty\}$ where $\mathcal{N}_\infty=\{F:D_\phi(F\|F_*)<\infty\}$, and $\vec{P}=(P,0_{d_3+d_4})$. For $A,B\subseteq\mathbb{R}^d$, we let $\mathrm{ri}(A)$ denote the relative interior of $A$ and $A+B=\{a+b:a\in A,b\in B\}$.

**Definition 2.1** Condition S *holds at* $(\theta,\gamma,P)$ *if* $\vec{P}\in\mathrm{ri}(\mathcal{G}(\theta,\gamma)+\mathcal{C})$.

Using relative interior instead of interior allows for moment functions that are collinear at some $\theta$ (i.e., some moments are redundant). To give some intuition, consider moment equality

[12] The $\mu$-essential supremum of a function $f$ is denoted $\mu$-$\mathrm{ess\,sup}\, f$ and is the smallest value $c$ for which $\mu(\{u:f(u)>c\})=0$. The $\mu$-essential infimum, denoted $\mu$-$\mathrm{ess\,inf}$, is defined analogously.

models. Condition S requires that (1) holds at $(\theta, \gamma, P)$ under some $F \in \mathcal{N}_\delta$ that is "interior" to $\mathcal{N}_\infty$, in the sense that one can perturb the (non-redundant) moments in any direction by perturbing $F$. For moment inequality models, Condition S also requires that there is $F \in \mathcal{N}_\infty$ under which all moment inequalities hold strictly at $(\theta, \gamma, P)$.

Let $\Theta_I = \{\theta \in \Theta : (1) \text{ holds for some } F \in \mathcal{F}_\theta\}$ denote the (nonparametric) identified set for $\theta$. Define the "nonparametric" objective function

$$\underline{K}_{np}(\theta; \gamma, P) = \inf_{F \in \mathcal{F}_\theta} \mathbb{E}^F[k(U, \theta, \gamma)] \quad \text{subject to (1) holding at } (\theta, \gamma, P)\,, \tag{18}$$

with the understanding that $\underline{K}_{np}(\theta; \gamma, P) = +\infty$ if the infimum runs over an empty set. Let $\overline{K}_{np}(\theta; \gamma, P)$ denote the analogous supremum. Evidently,

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

**Definition 2.2** *$\Theta_I$ is $S$-regular if for all $\epsilon > 0$ there exist $\underline{\theta}, \overline{\theta} \in \Theta_I$ such that Condition S holds at $(\underline{\theta}, \gamma_0, P_0)$ and $(\overline{\theta}, \gamma_0, P_0)$, $\underline{K}_{np}(\underline{\theta}; \gamma_0, P_0) < \inf \mathcal{K} + \epsilon$, and $\overline{K}_{np}(\overline{\theta}; \gamma_0, P_0) > \sup \mathcal{K} - \epsilon$.*

Intuitively, S-regularity requires that the values the counterfactual takes at "boundary" points of $\Theta_I$ (i.e., at which Condition S fails) are not materially more extreme than values it can take at points "inside" $\Theta_I$ (i.e., at which Condition S holds). This condition can be verified under more primitive continuity conditions on $k$ and $g$. A sufficient (but not necessary) condition for S-regularity is that Condition S holds at $(\theta, \gamma_0, P_0)$ for all $\theta \in \Theta_I$.

**Theorem 2.1** *Suppose that Assumption Φ holds, $k$ is $\mu$-essentially bounded, $\Theta_I$ is $S$-regular, and $\mu$ and $F_*$ are mutually absolutely continuous. Then*

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

Theorem 2.1 shows that our procedure can be used to approximate the sharp nonparametric bounds $\inf \mathcal{K}$ and $\sup \mathcal{K}$ by setting $\delta$ to be large but finite. If $\mu$ is Lebesgue measure—which it often is in applications—then the mutual absolute continuity condition in Theorem 2.1 is satisfied whenever $F_*$ has strictly positive density over $\mathcal{U}$.

**Remark 2.7** *Appendix B presents the dual forms of $\underline{K}_{np}$ and $\overline{K}_{np}$. Unlike $\underline{K}_\delta$ and $\overline{K}_\delta$, the duals of $\underline{K}_{np}$ and $\overline{K}_{np}$ are min-max and max-min problems which involve an inner optimization over $u$. These problems may be computationally challenging, especially when $u$ is multivariate. Comparing Proposition 2.1 with the duals in Appendix B, we see that setting $\delta < \infty$*

*replaces a "hard-max" (an optimization over $u$) with a "soft-max" (a convex expectation). In this respect, adding the constraint $F \in \mathcal{N}_\delta$ may be viewed as a regularization of the nonparametric objective functions, similar to the use of entropic penalization to regularize objective functions in optimal transport problems—see, e.g., Cuturi (2013).*

Theorem 2.1 is silent on the issue of how large $\delta$ needs to be so that $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are close to the nonparametric bounds. While this is model- and counterfactual-specific, the following toy example suggests that relatively small values of $\delta$ may suffice in some problems where the counterfactual is a choice probability.

**Example 2.4** Consider the problem

$$\overline{\kappa}_\delta = \sup_{\theta \in \mathbb{R}, F \in \mathcal{N}_\delta} \mathbb{E}^F[1\!\!1\{U \leq \theta\}] \quad \text{subject to} \quad \mathbb{E}^F[U - \theta] = 0,$$

where $\mathcal{N}_\delta$ is defined by KL divergence and $F_*$ is the $N(0,1)$ distribution. When $F = F_*$, the only solution to $\mathbb{E}^F[U - \theta] = 0$ is $\theta = 0$. Therefore, the value of the counterfactual under $F_*$ is $\mathbb{E}^{F_*}[1\!\!1\{U \leq 0\}] = \frac{1}{2}$ whereas $\sup \mathcal{K} = 1$. In Appendix H, we derive the large-$\delta$ approximation $\overline{\kappa}_\delta = 1 - 2\pi e^{-2\delta - 1}(1 + o(1))$. By symmetry, $\underline{\kappa}_\delta = 2\pi e^{-2\delta-1}(1 + o(1))$ and $\inf \mathcal{K} = 0$. Therefore, in this example, $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ converge rapidly to $\inf \mathcal{K}$ and $\sup \mathcal{K}$ as $\delta$ increases. □

More generally, suppose the dual problems (13) and (14) have unique solutions $\underline{\eta}$ and $\overline{\eta}$ for $\eta$ (where the optimization is performed over $\eta \geq 0$)[13] and the optimal $\underline{\eta}$ and $\overline{\eta}$ depend continuously on $\delta$. Under appropriate regularity conditions (see, e.g., Milgrom and Segal (2002)), it follows that

$$\frac{\partial \underline{K}_\delta(\theta; \gamma, P)}{\partial \delta} = -\underline{\eta}, \qquad \frac{\partial \overline{K}_\delta(\theta; \gamma, P)}{\partial \delta} = \overline{\eta}.$$

One can therefore infer from $\underline{\eta}$ and $\overline{\eta}$ the extent to which, if at all, the bounds at any fixed $\theta$ would widen further if $\delta$ was increased.

# 3 Practical Considerations

We now discuss practical details for implementing our procedure. Section 3.1 discusses computational methods, Section 3.2 presents our MPEC approach, and Section 3.3 discusses methods for dealing with over-identified models.

[13] Optimizing over $\eta \geq 0$ rather than $\eta > 0$ does not affect the optimal value—see Proposition G.1.

## 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 $\theta$.

The expectations in the objective functions (13), (14), and (16) are available in closed form for certain settings,[14] in which case the dimension of $u$ does not play a role in the computational complexity of our procedure. Otherwise, the expectations will need to be computed numerically. If so, the dimension of $u$ will play a role in terms of determining how many quadrature points or Monte Carlo draws are needed to control numerical approximation error. In the empirical applications we used a randomized quasi-Monte Carlo approach based on scrambled Halton sequences as in Owen (2017).

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 $g$ are differentiable in $\theta$.[15] Our procedures were all implemented in Julia with the inner and outer optimizations solved using KNITRO. A general-purpose implementation of our methods in Julia is provided in the supplemental material.

As with parameter estimation in nonlinear structural models, the outer optimization with respect to $\theta$ is typically non-convex. In applications, we iteratively applied a multi-start procedure in an attempt to converge to global optima. Computation times are reported in the applications below.

## 3.2 MPEC Approach

We now describe and formally justify an MPEC version of our procedure in the spirit of Su and Judd (2012). This approach simplifies computation in models with endogenous parameters defined by equilibrium conditions (e.g., value functions defined by Bellman equations), resulting in significant computational gains for DDC models in particular.

Suppose $\theta = (\theta_s, \theta_e)$ and $g_4 = (g_{4s}, g_{4e})$ where $\theta_s$ are "deep" structural parameters and $\theta_e$ are "endogenous" parameters that are defined implicitly by $g_{4e}$. That is, for any $(\theta_s, \gamma, F)$, the

[14] An earlier draft derived closed-form expressions for a discrete game of complete information with Gaussian payoff shocks and KL neighborhoods—see `https://arxiv.org/abs/1904.00989v2`.

[15] In practice, we smoothed any non-smooth moments and used automatic differentiation to compute derivatives with respect to $\theta$ if these were not easily available analytically.

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 $\theta_s = \theta_\pi$ or $(\theta_\pi, \beta)$, $\theta_e = (v, \tilde{v})$ collects the value functions in the baseline model and counterfactual, and $g_{4e}$ collects the functions representing the corresponding Bellman equations, as in display (6). While our procedure can be implemented as described in Section 2, this does not make use of the fact that $\theta_e$ is defined implicitly by $g_{4e}$.

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

$$\begin{aligned} &\mathbb{E}^F[g_1(U, \theta, \gamma_0)] \leq P_{10}, \quad \mathbb{E}^F[g_2(U, \theta, \gamma_0)] = P_{20}, \\ &\mathbb{E}^F[g_3(U, \theta, \gamma_0)] \leq 0, \qquad \mathbb{E}^F[g_{4s}(U, \theta, \gamma_0)] = 0, \end{aligned} \tag{19}$$

and define criterion functions using these only:

$$\underline{K}^s_\delta(\theta; \gamma, P) = \inf_{F \in \mathcal{N}_\delta} \mathbb{E}^F[k(U, \theta, \gamma)] \quad \text{subject to (19) holding at } (\theta, \gamma, P)\,, \tag{20}$$

$$\overline{K}^s_\delta(\theta; \gamma, P) = \sup_{F \in \mathcal{N}_\delta} \mathbb{E}^F[k(U, \theta, \gamma)] \quad \text{subject to (19) holding at } (\theta, \gamma, P)\,. \tag{21}$$

Under the conditions of Proposition 2.1, these criterion functions may be restated as

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

$$\overline{K}^s_\delta(\theta; \gamma, P) = \inf_{\eta > 0, \zeta \in \mathbb{R}, \lambda \in \Lambda_s} \eta \mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,\theta,\gamma) - \zeta - \lambda' g_s(U,\theta,\gamma)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda'_{12} P\,, \tag{23}$$

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 $d_{4s} = \dim(g_{4s})$. Problems (22) and (23) simplify analogously to (17) when $k$ does not depend on $u$, with the minimum divergence problem $\Delta$ defined using $g_s$ in place of $g$.

In our MPEC approach, the criterion functions (22) and (23) are optimized with respect to $\theta$, with the remaining moment conditions involving $g_{4e}$ appended as constraints. Importantly, these constraints are evaluated under the "least favorable" distributions $\underline{F}_{\delta,\theta}$ and $\overline{F}_{\delta,\theta}$ that solve problems (20) and (21), respectively. The following proposition formally justifies this approach.

**Proposition 3.1** *Suppose that Assumption $\Phi$ holds. Then the problems*

$$\inf_{\theta \in \Theta} \underline{K}_\delta(\theta; \gamma, P)$$

*and*

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

*have the same value. An analogous result holds for the upper bound.*

To implement our MPEC approach, note that the expectations in the constraints may be expressed in terms of changes of measure. Let $\underline{m}_{\delta,\theta} = \mathrm{d}\underline{F}_{\delta,\theta}/\mathrm{d}F_*$ and $\overline{m}_{\delta,\theta} = \mathrm{d}\overline{F}_{\delta,\theta}/\mathrm{d}F_*$ so that

$$\mathbb{E}^{\underline{F}_{\delta,\theta}}[\,\cdot\,] = \mathbb{E}^{F_*}[\underline{m}_{\delta,\theta}(U) \times \,\cdot\,]\,, \quad \mathbb{E}^{\overline{F}_{\delta,\theta}}[\,\cdot\,] = \mathbb{E}^{F_*}[\overline{m}_{\delta,\theta}(U) \times \,\cdot\,]\,.$$

If $k$ depends on $u$, then we construct $\underline{m}_{\delta,\theta}$ and $\overline{m}_{\delta,\theta}$ from solutions to (22) and (23), say $(\underline{\eta},\underline{\zeta},\underline{\lambda})$ and $(\overline{\eta},\overline{\zeta},\overline{\lambda})$ (these solutions exist under the regularity conditions below). If $\underline{\eta} > 0$, then the distribution solving (20) is unique and is induced by the change of measure

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

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 (24) by $(-\overline{\eta},-\overline{\zeta},-\overline{\lambda})$. The multiplier $\underline{\zeta}$ has a closed-form solution for KL neighborhoods, which yields

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

and similarly for $\underline{m}_{\delta,\theta}(u)$.

If $\underline{\eta} = 0$, then the constraint $D_\phi(F\|F_*) \leq \delta$ is slack and there may be multiple minimizing distributions. As shown in the proof of Proposition 3.2, each such distribution must be supported on

$$\underline{A}_{\delta,\theta} := \{u : k(u,\theta,\gamma) + \underline{\lambda}' g_s(u,\theta,\gamma) = F_*\text{-}\operatorname{ess\,inf}(k(\cdot,\theta,\gamma) + \underline{\lambda}' g_s(\cdot,\theta,\gamma))\}\,.$$

Note $F_*(\underline{A}_{\delta,\theta}) > 0$ is required for $\underline{\eta} = 0$ to be a solution. Otherwise, any distribution supported on $\underline{A}_{\delta,\theta}$ is not absolutely continuous with respect to $F_*$ and is therefore not in $\mathcal{N}_\delta$. If $\underline{\eta} = 0$ and $F_*(\underline{A}_{\delta,\theta}) > 0$, then we construct $\underline{m}_{\delta,\theta}$ by restricting $F_*$ to $\underline{A}_{\delta,\theta}$ and rescaling:

$$\underline{m}_{\delta,\theta}(u) = 1\!\!1\{u \in \underline{A}_{\delta,\theta}\}/F_*(\underline{A}_{\delta,\theta}).$$

The function $\overline{m}_{\delta,\theta}(u)$ is constructed analogously, replacing $\underline{\lambda}$ with $-\overline{\lambda}$ and the set $\underline{A}_{\delta,\theta}$ with $\overline{A}_{\delta,\theta} = \{u : k(u,\theta,\gamma) - \overline{\lambda}' g_s(u,\theta,\gamma) = F_*\text{-}\operatorname{ess\,sup}(k(\cdot,\theta,\gamma) - \overline{\lambda}' g_s(\cdot,\theta,\gamma))\}$.

If $k$ does not depend on $u$, then $\underline{m}_{\delta,\theta}$ and $\overline{m}_{\delta,\theta}$ are constructed from solutions to a version of problem (16) that uses only the reduced set of moments $g_s$ rather than $g$. Under the regularity conditions below, this program has a solution, say $(\underline{\zeta}, \underline{\lambda})$. In this case, we define

$$\underline{m}_{\delta,\theta}(u) = \overline{m}_{\delta,\theta}(u) = \dot{\phi}^\star\left(-\underline{\zeta} - \underline{\lambda}' g_s(u,\theta,\gamma)\right) . \tag{25}$$

For KL divergence the change of measure simplifies 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** *Suppose that Assumption* $\Phi$ *holds, Condition S holds at* $(\theta,\gamma,P)$*, and there exists a distribution* $F$ *with* $D(F\|F_*) < \delta$ *under which (19) holds at* $(\theta,\gamma,P)$*. Then the distributions* $\underline{F}_{\delta,\theta}$ *and* $\overline{F}_{\delta,\theta}$ *induced by* $\underline{m}_{\delta,\theta}$ *and* $\overline{m}_{\delta,\theta}$ *solve (20) and (21), respectively.*

**Example.** We consider a numerical example for the DDC model of Rust (1987) based on the parameterization in Section 5.4 of Norets and Tang (2014). The counterfactual they consider is a hypothetical change in the law of motion of the state. We follow these papers and use state-space of dimension 90. As $|\mathcal{S}| = 90$ and $\mathcal{D}_0 = \{0,1\}$, there are 90 functions in $g_2$ representing the observed CCPs. There are another 180 functions in $g_{4e}$ representing the Bellman equations in the baseline model and counterfactual across states. We also use two mean-zero normalizations $\mathbb{E}^F[U_d] = 0$ for $d \in \{0,1\}$, so $g_{4s}(U,\theta,\gamma) = (U_0, U_1)$. Our MPEC approach therefore uses 92 moments in the inner optimization (90 for CCPs and two mean-zero normalizations on the shocks) with the remaining 180 moments appended as constraints, while the full approach uses all 272 moments in the inner optimization.

Table 2 reports computation times for the inner optimization problems (14) and (23) (denoted $\overline{K}_\delta$) for maximizing the counterfactual CCP in the highest mileage state.[16] We also report times for solving the minimum divergence problem (16) (denoted $\Delta$) using the full set of moment functions $g$ and its MPEC analogue using $g_s$. Neighborhoods are constrained by a hybrid of KL and $\chi^2$ divergence as in the empirical applications—see Section 5. As can be seen, the inner optimization problems are solved at least 20 times faster for the MPEC implementation, with the relative efficiency increasing in $\delta$.

[16] The times in Table 2 are based on initializing the solver at $\eta = 1$, $\zeta = 0$, and $\lambda = 0$. When embedded in the outer optimization over $\theta$, computation times for the inner problem are reduced significantly by using a warm start that initializes at the solution to the inner problem at the previous value of $\theta$.

Table 2: Computation times (in seconds) for the inner problems

| Implementation | Objective | | | |
|---|---|---|---|---|
| | $\overline{K}_{0.01}$ | $\overline{K}_{0.10}$ | $\overline{K}_{1.00}$ | $\Delta$ |
| MPEC (92 moments) | 0.207 | 0.232 | 0.256 | 0.108 |
| Full (272 moments) | 4.317 | 12.978 | 43.699 | 3.365 |

*Note:* Expectations are computed using 50,000 scrambled Halton draws. Computations are performed in Julia v1.6.4 and KNITRO v12.4.0 on a 2.7GHz MacBook Pro with 16GB memory.

## 3.3 Over-identification

In over-identified models (i.e., where the number of moment conditions $d$ exceeds the dimension $d_\theta$ of $\theta$), there might not exist $\theta \in \Theta$ for which the sample moment conditions (7) hold under $F_*$. We propose two methods for handling over-identified models.

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

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

The interval $[\hat{\underline{\kappa}}_\delta, \hat{\overline{\kappa}}_\delta]$ will be nonempty for $\delta > \hat{\delta}$. If the model is correctly specified under $F_*$,[17] then $\hat{\delta}$ will converge in probability to zero under the conditions of Theorem 6.1. In this case, the interval $[\hat{\underline{\kappa}}_\delta, \hat{\overline{\kappa}}_\delta]$ will be nonempty with probability approaching one for each fixed $\delta > 0$.

It is also possible that $\hat{\delta} = +\infty$ in correctly specified but over-identified models when $\hat{P}$ is incompatible with certain model restrictions. For instance, CCPs are often estimated non-parametrically using empirical choice frequencies. If some choices aren't observed in the data, then the estimated CCPs will be zero even though model-implied CCPs are strictly positive.

This issue can be circumvented in models defined by equality restrictions only (hence $P_0 \equiv P_{20}$) using the following two-step approach. First, compute a preliminary estimator $\tilde{\theta}$ of $\theta$ based on (7). Then, set $\hat{P} = \mathbb{E}^{F_*}[g_2(U, \tilde{\theta}, \hat{\gamma})]$. This second-step estimator $\hat{P}$ is compatible with the model by construction, thereby ensuring that the interval $[\hat{\underline{\kappa}}_\delta, \hat{\overline{\kappa}}_\delta]$ is nonempty for each $\delta > 0$. The estimator $\hat{P}$ will be consistent and asymptotically normal under mild regularity conditions provided the model is correctly specified under $F_*$, so the consistency and inference results developed in Section 6 will also apply.

[17] Neither our theoretical results developed in Section 2 nor the estimation and inference results in Section 6 require correct specification of the model under $F_*$.

# 4 Interpreting the Neighborhood Size

This section presents some theoretical results and practical methods to help interpret the neighborhood size $\delta$. Sections 4.1 and 4.4 discuss properties of $\phi$-divergences and their implications for interpreting $\delta$. Section 4.2 shows how to construct the "least favorable" distributions that minimize or maximize the counterfactual. Section 4.3 gives a practical, model-based metric for interpreting $\delta$.

## 4.1 Invariance

A defining property of $\phi$-divergences are their invariance to invertible transformations. That is, if $T$ is an invertible transformation and $G$ and $G_*$ denote the distributions of $T(U)$ when $U \sim F$ and $U \sim F_*$, respectively, then $D_\phi(F \| F_*) = D_\phi(G \| G_*)$.[18] An important consequence of invariance is that $\delta$ has the same interpretation under a change in units. For instance, if one researcher writes a model in terms of dollars with $U \sim F_*$ and another researcher uses thousands of dollars with $U \sim G_*$ for $G_*(u) = F_*(10^{-3}u)$, then $F$ is in $\mathcal{N}_\delta$ if and only if its rescaled counterpart $G$ is in a $\delta$-neighborhood of $G_*$. A second consequence is that neighborhood size is invariant under invertible location and scale transformations of $F_*$ (e.g., $N(\mu, \Sigma)$ versus $N(0, I)$).

## 4.2 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 may easily be recovered. To help interpret $\delta$, one may plot the LFDs and compute other quantities of interest (e.g., correlations or welfare measures) under them.

Section 3.2 describes how to construct LFDs when our MPEC approach is used. LFDs for our full (i.e., non-MPEC) approach are a special case with $g_4 = g_{4s}$. To briefly summarize, consider the LFD $\underline{F}_{\delta,\theta}$ solving the minimization problem (11). First suppose that $k$ depends on $u$. Let $(\underline{\eta}, \underline{\zeta}, \underline{\lambda})$ solve problem (13). If $\underline{\eta} > 0$, then $\underline{F}_{\delta,\theta}$ is unique and its change-of-measure $\underline{m}_{\delta,\theta} = \mathrm{d}\underline{F}_{\delta,\theta} / \mathrm{d}F_*$ is given by

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

[18] See, e.g., Liese and Vajda (1987). A more direct statement is in Qiao and Minematsu (2010).

The LFD $\overline{F}_{\delta,\theta}$ solving the maximization problem (12) is constructed similarly, replacing $(\underline{\eta}, \underline{\zeta}, \underline{\lambda})$ in (26) with $(-\overline{\eta}, -\overline{\zeta}, -\overline{\lambda})$, where $(\overline{\eta}, \overline{\zeta}, \overline{\lambda})$ solves (14). If $\underline{\eta} = 0$ or $\overline{\eta} = 0$, then there may exist multiple distributions solving (11) and (12) at $\theta$. LFDs in this case are constructed analogously to the method described in Section 3.2. Note that $\underline{\eta} = 0$ or $\overline{\eta} = 0$ is unlikely if $k$ and/or elements of $g$ are unbounded in $u$—see the discussion in Section 3.2. If $k$ does not depend on $u$, then we set

$$\underline{m}_{\delta,\theta}(u) = \overline{m}_{\delta,\theta}(u) = \dot{\phi}^{\star}\left(-\underline{\zeta} - \underline{\lambda}' g(u, \theta, \gamma)\right) \tag{27}$$

where $(\underline{\zeta}, \underline{\lambda})$ solves (16). While there may exist multiple distributions solving (11) and (12) in this case, the distribution induced by (27) has smallest $\phi$-divergence relative to $F_*$.

## 4.3 Viewing Neighborhood Size through the Lens of the Model

Another method for interpreting $\delta$ is based on measuring the variation in the moments at the distributions solving (8) and (9) relative to their values under $F_*$.

Consider the sets of minimizing and maximizing values of $\theta$ at which $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are attained, say $\underline{\Theta}_\delta$ and $\overline{\Theta}_\delta$. These are nonempty under the regularity conditions in Section 6. While the moment conditions (1) hold at any $\theta \in \underline{\Theta}_\delta \cup \overline{\Theta}_\delta$ under the corresponding LFD, they will typically not hold at $\theta$ under $F_*$. We therefore define

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

where $(v)_+ = (\max\{v_i, 0\})_{i=1}^d$ for a vector $v \in \mathbb{R}^d$. The quantity $size(\delta)$ is the maximum degree to which the moments at $\theta \in \underline{\Theta}_\delta \cup \overline{\Theta}_\delta$ violate (1) under $F_*$.

This measure is informative about the extent to which the distortions to $F_*$ required to attain the smallest and largest values of the counterfactual over $\mathcal{N}_\delta$ are reflected in (1). Small values of $size(\delta)$ indicate that 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. In practice, this measure can be computed by replacing $(P_0, \gamma_0)$ by estimators $(\hat{P}, \hat{\gamma})$ and $\underline{\Theta}_\delta$ and $\overline{\Theta}_\delta$ by the minimizers and maximizers of the sample criterions or by the estimators of $\underline{\Theta}_\delta$ and $\overline{\Theta}_\delta$ introduced in Section 6.

### 4.4 Relating Different Divergences

It is well known that $\phi$-divergences are equivalent over local neighborhoods (see, e.g., Theorem 4.1 of Csiszár and Shields (2004)). However, the bounds $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ may depend on the choice of $\phi$ when $\delta$ is not arbitrarily small. In this case, bounds induced by different $\phi$ functions may be related as follows. Let $\mathcal{N}_{\delta,1}$ and $\mathcal{N}_{\delta,2}$ denote $\delta$-neighborhoods induced by $\phi_1$ and $\phi_2$, respectively. The quantity

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

is a measure of relative neighborhood size: if $\bar{a} < \infty$ then $\mathcal{N}_{\delta,2} \subseteq \mathcal{N}_{\bar{a}\delta,1}$ for each $\delta > 0$ (see the proof of Proposition 4.1). For instance, we obtain $\bar{a} = 2$ when comparing KL divergence ($\phi_1(x) = x \log x - x + 1$) and $\chi^2$ divergence ($\phi_2(x) = \frac{1}{2}(x-1)^2$). Therefore, $\delta$-neighborhoods under $\chi^2$ divergence are contained in $2\delta$-neighborhoods under KL divergence. Interchanging $\phi_1$ and $\phi_2$ produces $\bar{a} = +\infty$, which reflects the fact that KL divergence is weaker than $\chi^2$ divergence.

Let $\underline{\kappa}_{\delta,1}$ and $\underline{\kappa}_{\delta,2}$ denote the smallest counterfactual from display (8) over $\mathcal{N}_{\delta,1}$ and $\mathcal{N}_{\delta,2}$, respectively. Define $\overline{\kappa}_{\delta,1}$ and $\overline{\kappa}_{\delta,2}$ analogously.

**Proposition 4.1** *Suppose that Assumption $\Phi$ holds for both $\phi_1$ and $\phi_2$ and $\bar{a}$ is finite. Then $[\underline{\kappa}_{\delta,2}, \overline{\kappa}_{\delta,2}] \subseteq [\underline{\kappa}_{\bar{a}\delta,1}, \overline{\kappa}_{\bar{a}\delta,1}]$ for each $\delta > 0$.*

It follows from Proposition 4.1 that bounds that are wide under $\phi_2$ must necessarily be wide under $\phi_1$. Similarly, narrow bounds under $\phi_1$ must also be narrow under $\phi_2$. Note also that the inclusion in Proposition 4.1 holds for any counterfactual.

## 5 Empirical Applications

### 5.1 Marital College Premium

Chiappori et al. (2017), henceforth CSW, study the evolution of marital returns to education using a frictionless matching model with transferable utility following (Choo and Siow, 2006). Within this framework, the "marital college premium" is the additional expected utility that an individual would derive from the marriage market if they had a (counterfactually) higher level of education. CSW find that marital college premiums for women in the United States increased significantly across cohorts from the mid to late 20th century, particularly for the more highly educated.

As is conventional following Dagsvik (2000) and Choo and Siow (2006), CSW assume latent variables representing individuals' idiosyncratic marital preferences are i.i.d. Gumbel. The marital college premium is only partially identified when the distribution of these latent variables is not specified. We therefore perform a sensitivity analysis of CSW's estimates to departures from this conventional parametric assumption.

Our analysis makes several findings. First, it seems impossible to draw conclusions about how the marital college premium has changed over time under small nonparametric relaxations of the i.i.d. Gumbel assumption. Premiums have narrow nonparametric bounds at fixed parameter values, but slight relaxations of the i.i.d. Gumbel assumption allows for significant variation in parameters which, in turn, produces uninformatively wide bounds. As parameters are just-identified under any fixed distribution of shocks (Galichon and Salanié, 2021), further restrictions on parameters or shape restrictions on the distribution are required to tighten the bounds. We show that imposing exchangeability can tighten the bounds significantly.

**Model and Benchmark Estimates.** Agents are male or female and one of $J$ types (education levels). A type-$a$ male receives utility $\varepsilon_{a0}$ if he chooses to be unmatched and $z_{ab} + \varepsilon_{ab}$ if he matches with a type-$b$ female. Similarly, a type-$b$ female receives utility $e_{0b}$ if she chooses to be unmatched and $t_{ab} + e_{ab}$ 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. The latent shocks $(\varepsilon_{a0}, \ldots, \varepsilon_{aJ})$ and $(e_{0b}, \ldots, e_{Jb})$ represent individuals' idiosyncratic marital preferences. Shocks are i.i.d. across individuals and have mean zero. The type $b$ to $b'$ marital education premium for females is the difference in expected marital utility

$$\kappa = \mathbb{E}^F \left[ \max_{a=0,\ldots,J} \Big( t_{ab'} + e_{ab'} \Big) \right] - \mathbb{E}^F \left[ \max_{a=0,\ldots,J} \Big( t_{ab} + e_{ab} \Big) \right], \tag{28}$$

where $F$ denotes the distribution of $(e_{0b}, \ldots, e_{Jb'})$ and $t_{0b} = t_{0b'} = 0$.

CSW use data from the American Community Survey. They form 28 cohorts indexed by female birth year from 1941 (cohort 1) to 1968 (cohort 28), each of which is treated as an independent marriage market. We focus on CSW's estimates for whites. There are $J = 5$ types: "high-school dropouts", "high-school graduates", "some college", "college graduate", and "college-plus". We center our analysis on the "some college" to "college graduate" premium. Figure 1 presents estimates and 95% confidence sets (CSs) for the premium under the i.i.d. Gumbel assumption (cf. Figure 21 in CSW) based on CSW's replication files.

**Implementation.** The model reduces to a standard individual-level discrete choice problem for each type (see CSW's Propositions 1 and 2). We assume that the distribution of females' preference shocks does not depend on their type, so we drop the $b$ subscript and consider a single random vector $U = (e_0, \ldots, e_J)$. We allow the distribution $F$ of $U$ to vary across cohorts and implement our procedures cohort-by-cohort.[19]

Under any fixed $F$, a cohort's parameters $(t_{ab})_{a=1}^{J}$ are just-identified from the marriage probabilities for that cohort's type-$b$ women (Galichon and Salanié, 2021). We therefore impose only the moment conditions involving the parameters $\theta = (t_{ab}, t_{ab'})_{a=1}^{J}$ appearing in (28), as the remaining parameters can be chosen to fit the remaining marriage probabilities under the resulting least-favorable distribution. We form $g_2$ to explain the type $b$ and $b'$ marriage probabilities for women in a given cohort:

$$g_2(U,\theta) = \begin{bmatrix} (1\!\!1\{t_{ab} + e_a = \max_{a'=0,\ldots,J}(t_{a'b} + e_{a'})\})_{a=1}^{J} \\ (1\!\!1\{t_{ab'} + e_a = \max_{a'=0,\ldots,J}(t_{a'b'} + e_{a'})\})_{a=1}^{J} \end{bmatrix}$$

and form $\hat{P}_2$ using CSW's estimates of the corresponding type-$b$ and $b'$ marriage probabilities. We set $g_4(U,\theta) = (e_j, e_j^2 - \pi^2/6)_{j=0}^{J}$ so that shocks have mean zero and the same variance as the Gumbel distribution. The scale normalization also ensures that the nonparametric bounds on the premium are finite at any fixed $\theta$. As $J = 5$, there are 22 moments (10 for marriage probabilities and 12 location/scale normalizations), and $\theta$ has dimension $10$.

We consider a second implementation which imposes invariance of $F$ under rotations and reflections of potential spouse types, so that the model-implied marriage probabilities depend on $\theta$ but not the labeling of potential spouse types (though they may depend on their ordering).[20] Formally, this shape restriction corresponds to dihedral exchangeability (see Appendix A.1); we refer to it simply as "exchangeability". Under this shape restriction, $F$ must satisfy the 22 moment conditions under all 12 rotations and reflections of the elements of $U$. This implementation therefore imposes a total of $264$ moment conditions. Rather than including all 264 moments separately, it suffices to form $g_2$ and $g_4$ by taking the averages of the 22 moments across the 12 permutations (see Appendix A.1). Both implementations therefore have inner optimization problems of the same dimension.

Computations are performed as described in Section 3.1. The first implementation uses 50,000 scrambled Halton draws to compute the expectations. The second uses 10,000 draws which are concatenated over the 12 permutations (see Remark A.2), for a total of 120,000 draws. Computation times are reported in Appendix D.1. CSs for $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are computed using

[19] In view of the just-identification results of Galichon and Salanié (2021), we would obtain the same bounds if $F$ was homogeneous across cohorts. Allowing for heterogeneity in own-type would result in wider bounds.

[20] Allowing dependence on the ordering of types seems desirable here as types correspond to education levels, which are naturally ordered.

Table 3: Metrics for interpreting $\delta$

| | Without exchangeability | | | With exchangeability | | |
|---|---|---|---|---|---|---|
| $\delta$ | $\rho_{\max}$, $\underline{\kappa}_\delta$ | $\rho_{\max}$, $\overline{\kappa}_\delta$ | $size$ | $\rho_{\max}$, $\underline{\kappa}_\delta$ | $\rho_{\max}$, $\overline{\kappa}_\delta$ | $size$ |
| 0.01 | -0.015 | -0.014 | 0.010 | -0.022 | 0.013 | 0.006 |
| 0.10 | -0.071 | -0.073 | 0.038 | -0.061 | 0.054 | 0.023 |
| 1 | -0.247 | -0.197 | 0.112 | -0.139 | 0.115 | 0.099 |
| 10 | -0.502 | -0.496 | 0.242 | -0.204 | 0.236 | 0.176 |
| 100 | -0.620 | -0.576 | 0.266 | -0.266 | 0.284 | 0.178 |

*Note:* Averages across cohorts of the largest element of the correlation matrix for $U$ under the LFDs at which the estimated lower bounds ($\rho_{\max}$, $\underline{\kappa}_\delta$) and upper bounds ($\rho_{\max}$, $\overline{\kappa}_\delta$) are attained, and our $size$ measure from Section 4.3. Each is computed at the parameter values at which the estimated upper and lower bounds are attained.

the bootstrap procedure in Section 6.2. Appendix D.1 discusses bootstrap implementation details and presents projection CSs using the method from Section 6.3.

We define neighborhoods using a hybrid of KL and $\chi^2$ divergence:

$$\phi(x) = \left[ \begin{array}{ll} x \log x - x + 1 & \text{if } x \leq e, \\ \frac{1}{2e}(x-e)^2 + (x-e) + 1 & \text{if } x > e. \end{array} \right.$$

We use this divergence because Assumption $\Phi$(ii) fails for KL divergence, whereas hybrid divergence only requires finite second moments for Assumption $\Phi$(ii). The LFDs under hybrid divergence are also everywhere positive, which is not guaranteed under $\chi^2$ divergence. We repeated our analysis with neighborhoods constrained by $\chi^2$ and $L^4$ divergences as robustness checks. Overall, our findings are not sensitive to $\phi$ (see Appendix D.1 for a discussion).

**Findings.** Figure 1 presents a sensitivity analysis of the "some college" to "college graduate" premium. Cohort-wise estimates and CSs for $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are presented, beginning at $\delta = 0.01$ and increasing $\delta$ by factors of 10 up to $\delta = 100$. Even with $\delta = 0.01$, estimates of $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ lie uniformly below and above zero across cohorts without exchangeability (see Figure 1a). Without further restrictions, it seems impossible to say whether the premium has increased or decreased over time under slight nonparametric relaxations of the i.i.d. Gumbel assumption. Figure 1b shows that imposing exchangeability can tighten the bounds, with the bounds for $\delta = 0.01$ significantly negative in early cohorts and significantly positive in later cohorts. But the $\delta = 0.1$ bounds with exchangeability again contain zero across all cohorts. Bounds for larger $\delta$ presented in Figures 1c and 1d are uninformatively wide.

To understand better what is meant by "small" and "large" neighborhoods, Figure 2 plots marginal CDFs for the LFDs under which the upper bounds for cohort 1 are attained. Similar

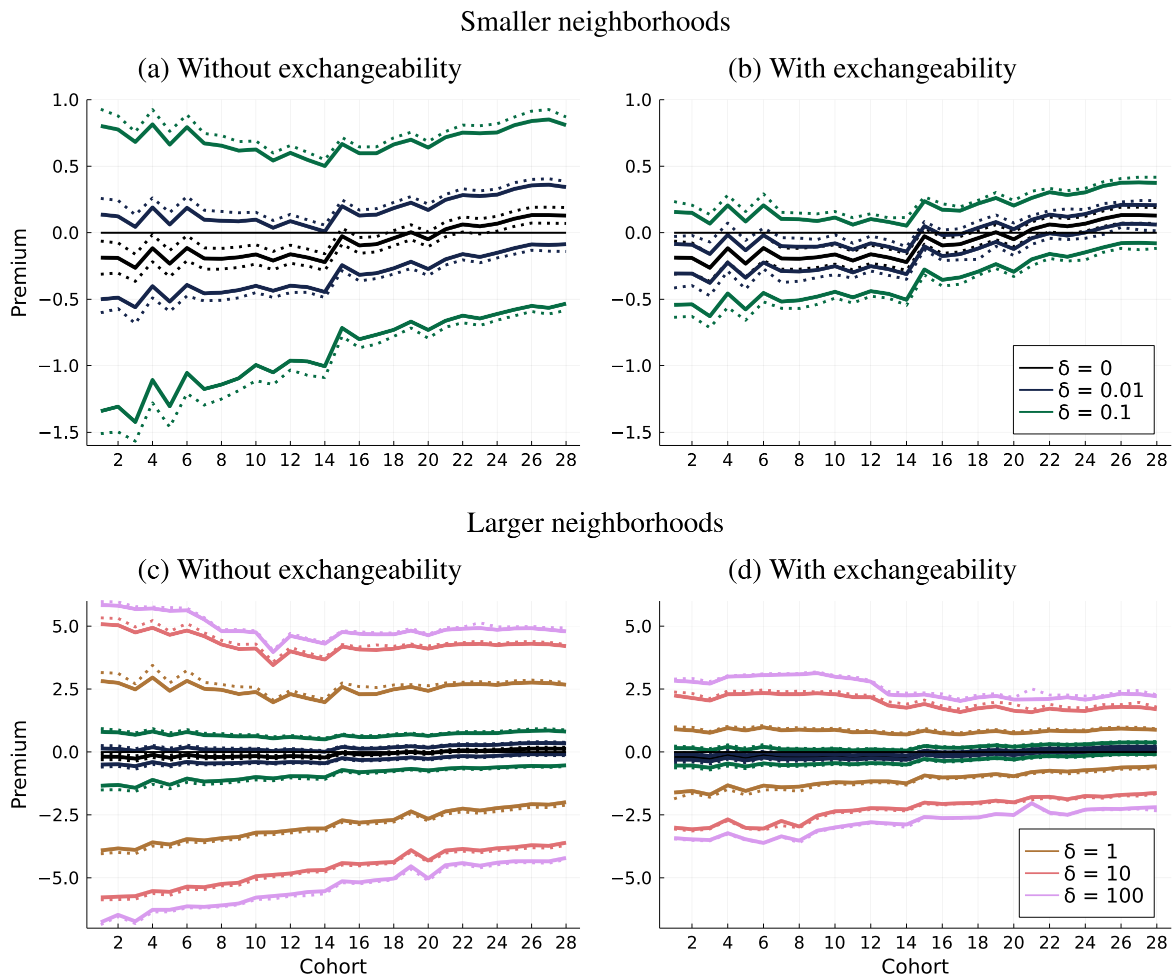


Figure 1: Sensitivity analysis of the "some college" to "college graduate" premium across cohorts. *Note:* Solid lines are estimates, dotted lines are (cohort-wise) 95% CSs. CSW's estimates and CSs correspond to $\delta = 0$.

LFDs (not reported) were obtained for other cohorts and the lower bounds. Without exchangeability, the LFDs with $\delta = 0.1$ are almost identical to Gumbel (plots with $\delta = 0.01$ are indistinguishable from Gumbel). LFDs appear close to Gumbel across most potential spouse types with $\delta = 1$, while for $\delta = 10$ and $\delta = 100$ the LFDs have kinks and indicate shifts in mass from the center of the distribution to the tails.

Under exchangeability (Figure 2b), the marginal distribution of shocks is independent of potential spouse type. In this case the LFDs for $\delta = 1$ or smaller are virtually indistinguishable from Gumbel. LFDs with $\delta = 10$ and $\delta = 100$ are also less kinked than Figure 2a because distortions are spread more evenly across potential spouse types.

We also computed the largest correlation of shocks under the LFDs at which the bounds are

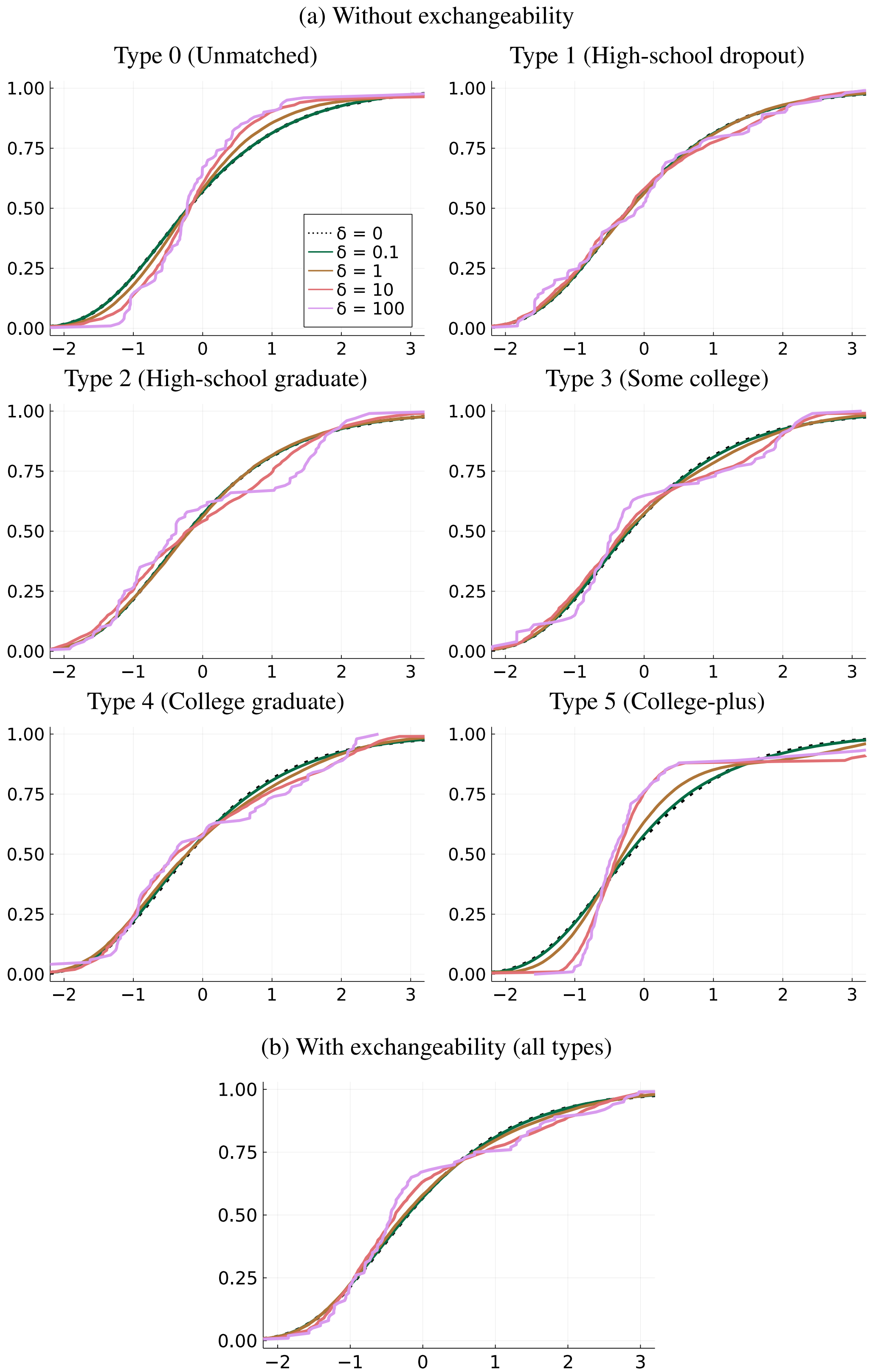


Figure 2: Marginal CDFs for the LFDs maximizing the "some college" to "college graduate" premium in cohort 1 across potential spouse types.

attained and our $size$ measure from Section 4.3. As these quantities are stable across cohorts, we present their averages in Table 3. Shocks are independent when $\delta = 0$ and only very weakly correlated for small $\delta$, while for large $\delta$ some shocks are strongly negatively correlated. The maximal correlations under exchangeability are smaller, especially for large $\delta$. Turning to the $size$ measure, the LFDs for $\delta = 0.01$ without exchangeability shift the model-implied marriage probabilities by 0.01 (on average, across cohorts) from their values under the i.i.d. Gumbel assumption. LFDs for $\delta = 10$ and $\delta = 100$ shift marriage probabilities around 0.25 (on average, across cohorts). Imposing exchangeability reduces the $size$ measure by around 25% because model parameters do not vary as much under this shape restriction.

In view of the small-$\delta$ bounds in Figure 1, the LFDs in Figure 2, and the metrics in Table 3, it seems impossible to draw conclusions about how the sign of the premium has changed over time under slight nonparametric relaxations of the i.i.d. Gumbel assumption. To help understand why, Figure 5 plots bounds where $F$ is allowed to vary but $\theta$ is held fixed at CSW's estimates. These "fixed-$\theta$" bounds for $\delta = 10$ and $\delta = 100$ are almost identical, and are roughly the same width as the $\delta = 0.01$ bounds in Figure 1. The width of the bounds in Figure 1 therefore seems largely due to the additional variation in $\theta$ that is permitted when parametric assumptions for $F$ are relaxed.

Overall, our findings are complementary to Gualdani and Sinha (2020) who perform a non-parametric reanalysis of CSW using the PIES methodology of Torgovitsky (2019b). Although they do not derive nonparametric bounds on the marital education premium itself, only terms that contribute to it, they also find no evidence of an increase in premiums across cohorts.

## 5.2 Welfare Analysis in a Rust Model

Our second empirical illustration is a sensitivity analysis for welfare counterfactuals in the DDC model of Rust (1987).

**Model and Benchmark Estimates.** We focus on the specification in Table IX of Rust (1987) where maintenance costs are linear in the state (i.e., mileage). In the notation of Example 2.3, $|\mathcal{S}| = 90$, $\beta = 0.9999$, and $\theta_\pi = (RC, MC)$ where $RC$ is the replacement cost and $MC$ is a maintenance cost parameter. Our counterfactual of interest is the change in average welfare arising from a 10% reduction in maintenance costs. Hence, $\pi_{1,s}(\theta_\pi) = \tilde{\pi}_{1,s}(\theta_\pi) = -RC$ and $\pi_{0,s}(\theta_\pi) = -0.001MC \times s$ (baseline) and $\tilde{\pi}_{0,s}(\theta_\pi) = 0.9\pi_{0,s}(\theta_\pi)$ (counterfactual). The counterfactual function is $k(\theta, \gamma) = w'(\tilde{v} - v)$ where $w$ is the stationary distribution of the state in the baseline model.

Under the i.i.d. Gumbel assumption, the estimated counterfactual at the maximum likelihood estimate (MLE) of $\theta_\pi$ is 73.07 and its 95% CS is [48.25,101.31].[21] The counterfactual is point-identified under the i.i.d. Gumbel assumption because $\theta_\pi$ is point-identified.

**Implementation.** We estimate CCPs using Rust's Group 4 data. Nonparametric estimates of the 90 CCPs are zero in many states, so we proceed as in Section 3.3 and take the model-implied CCPs at the MLE of $\theta_\pi$ (under the i.i.d. Gumbel assumption) as our estimate $\hat{P}_2$. We drop moment conditions for CCPs in states where the replacement probability is less than $0.001$ to avoid numerical instabilities induced by including near-degenerate moments. This reduces the dimension of $g_2$ to 71. We normalize $F$ so that shocks have mean zero and the same variance as the Gumbel distribution by appending $\mathbb{E}^F[U_d] = 0$ and $\mathbb{E}^F[U_d^2 - \pi^2/6] = 0$, for $d = 0, 1$, to $g_4$. 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 implement our methods as described in Section 3.2. The inner optimization uses 75 moments (71 for CCPs and 4 for normalizations), with the remaining 180 moments appended as constraints in the outer optimization. We define neighborhoods using hybrid divergence from Section 5.1 so that Assumption $\Phi$(ii) holds. Similar results are obtained with $\chi^2$ and $L^4$ neighborhoods (see Appendix D.2). Expectations are computed using 50,000 scrambled Halton draws—see Appendix D.2 for computation times. One-sided 95% CSs for $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are computed using the bootstrap procedure in Section 6.2. Appendix D.2 discusses bootstrap implementation details and presents projection CSs using the method from Section 6.3.

**Findings.** Estimates and confidence sets for $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$ are plotted in Figure 3 for values of $\delta$ from $0.01$ to $100$.[22] As can be seen, the bounds expand rapidly under slight relaxations of the i.i.d. Gumbel assumption then stabilize at around $\delta = 1$, where the lower bound is 6.45 and the upper bound is 160.5 which represents approximately 220% of the value under the i.i.d. Gumbel assumption.

To interpret $\delta$, in Figure 4 we plot the CDFs of $U_1 - U_0$ under the LFDs at which the estimated bounds $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained. LFDs were computed as described in Section 4.2 using the construction (27). The distributions appear very close to logistic (their distribution

[21] To construct this CS, we draw $\hat{\theta}_\pi^* \sim N(\hat{\theta}_\pi, \hat{\Sigma})$ where $\hat{\theta}_\pi$ is the MLE and $\hat{\Sigma}$ is an estimate of the inverse information matrix. For each $\hat{\theta}_\pi^*$ draw, we compute the baseline and counterfactual value functions $v^*$ and $\tilde{v}^*$, and hence the counterfactual $\hat{\kappa}^* = w'(\tilde{v}^* - v^*)$.

[22] The width of the CSs relative to the bounds reduces as $\delta$ gets large. We re-estimated our bounds using several different draws of bootstrapped CCPs in place of $\hat{P}_2$ and obtained bounds that spanned a range similar to our CSs for small $\delta$, but which for many draws converged to our estimates as $\delta$ increased. This corroborates the behavior of our bootstrap CSs. We conjecture that ultimately other features of the model are more important than the numerical values of the CCPs in determining nonparametric bounds on the welfare counterfactual.

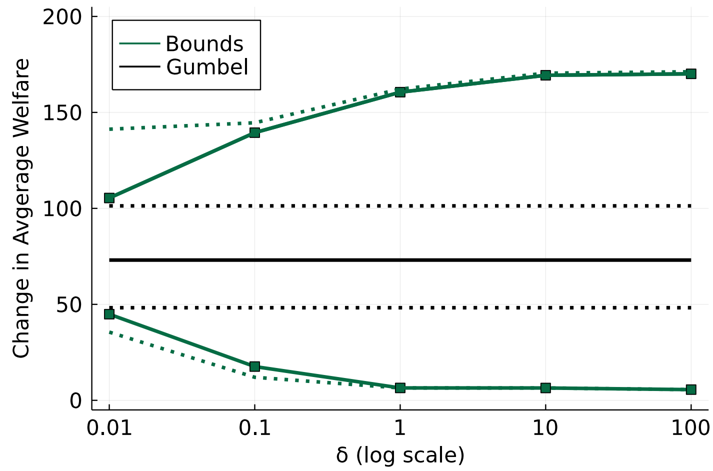


Figure 3: Sensitivity analysis for change in average welfare under a 10% maintenance cost subsidy. *Note:* Solid lines are estimates, dotted lines are 95% CSs.

when $\delta = 0$) for $\delta = 0.01$. Therefore, we see that large differences in welfare counterfactuals can arise under very slight departures from the i.i.d. Gumbel assumption. LFDs for the upper bound shift increasing amounts of mass to the center of the distribution of $U_1 - U_0$ as $\delta$ increases. LFDs corresponding to the lower bound are relatively less distorted, but have increasing amounts of mass shifted into the right tail. These are similar for $\delta = 0.1$ through $\delta = 100$ because the estimated lower bound stabilizes for smaller values of $\delta$ than the upper bound (cf. Figure 3).

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 $\underline{\hat{\kappa}}_\delta$ and $\hat{\bar{\kappa}}_\delta$ are attained. These are very small for $\delta = 0.01$ and remain small under the LFDs for $\underline{\hat{\kappa}}_\delta$ as $\delta$ increases, while $U_0$ and $U_1$ are strongly positively correlated under the LFDs for $\hat{\bar{\kappa}}_\delta$, especially for larger $\delta$ values. Given the asymmetry in distortions between the lower and upper values, we compute our $size$ measure separately for both. We measure distortions 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 0.016. By contrast, the LFDs for $\delta = 10$ and $\delta = 100$ shift the model-implied CCPs from their values under the i.i.d. Gumbel assumption by at most $0.04$ for $\underline{\hat{\kappa}}_\delta$ and $0.46$ for $\hat{\bar{\kappa}}_\delta$.

The parameters at which $\underline{\hat{\kappa}}_\delta$ and $\hat{\bar{\kappa}}_\delta$ are attained are also revealing about neighborhood size. 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 $\hat{\bar{\kappa}}_\delta$ are attained at very different parameter values, with much smaller cost parameters for the lower bound and larger parameters for the

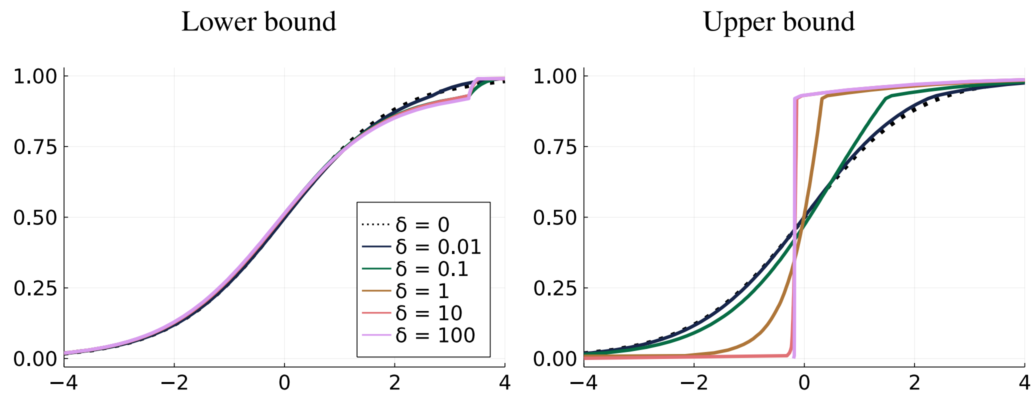


Figure 4: CDFs of $U_1 - U_0$ under the LFDs at which the estimated lower and upper bounds on the welfare counterfactual are attained.

Table 4: Metrics for interpreting $\delta$

| | Lower bound | | | | Upper bound | | | |
|---|---|---|---|---|---|---|---|---|
| $\delta$ | $corr$ | $size$ | $RC$ | $MC$ | $corr$ | $size$ | $RC$ | $MC$ |
| 0 | 0.000 | 0.000 | 10.208 | 2.294 | 0.000 | 0.000 | 10.208 | 2.294 |
| 0.01 | 0.036 | 0.010 | 7.357 | 1.411 | -0.027 | 0.016 | 13.390 | 3.307 |
| 0.1 | -0.058 | 0.039 | 5.186 | 0.553 | 0.149 | 0.109 | 16.134 | 4.374 |
| 1 | -0.045 | 0.039 | 4.023 | 0.203 | 0.616 | 0.346 | 17.166 | 5.038 |
| 10 | -0.040 | 0.039 | 4.022 | 0.202 | 0.765 | 0.461 | 17.595 | 5.331 |
| 100 | -0.063 | 0.039 | 3.931 | 0.176 | 0.764 | 0.469 | 17.626 | 5.365 |

*Note:* Correlation of $U_0$ and $U_1$ under the LFD at which the estimated lower and upper bounds are attained ($corr$), our $size$ measure from Section 4.3, and replacement and maintenance cost parameters at which the estimated lower and upper bounds are attained.

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. Correspondingly, a low $RC$ is needed to help the model to fit the observed CCPs at the smaller $MC$. While it is known that payoff parameters are not identified without parametric assumptions on $F$, it is perhaps surprising that these parameters vary by so much under slight relaxations of the i.i.d. Gumbel assumption. For instance, with $\delta = 0.01$ the lower bound is attained with cost parameters $RC = 7.357$ and $MC = 1.411$ while the upper bound is attained with cost parameters that are roughly double these values.

# 6 Estimation and Inference

We begin in Section 6.1 by establishing consistency and the asymptotic distribution of the estimators $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ from Section 2.4. We then present a bootstrap-based inference method in Section 6.2 and a projection-based inference method in Section 6.3.

## 6.1 Large-sample Properties of Plug-in Estimators

We first introduce some regularity conditions. Recall the space $\mathcal{E}$ from Assumption $\Phi$. We equip $\mathcal{E}$ with the Orlicz norm (see Appendix F)

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

This norm is equivalent to the $L^2(F_*)$ norm for $\chi^2$ and hybrid divergence and equivalent to the $L^q(F_*)$ norm for $L^p$ divergence ($p^{-1} + q^{-1} = 1$), while for KL divergence it is stronger than any $L^p(F_*)$ norm with $p < \infty$ but weaker than the sup-norm. 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 also require a slightly stronger notion of constraint qualification than Condition S from Section 2.5.

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

Condition S' replaces "relative interior" in Condition S with "interior". Finally, recall $\Delta(\theta; \gamma, P)$ from (16) and let $\Theta_\delta(\gamma, P) = \{\theta \in \Theta : \Delta(\theta; \gamma, P) < \delta\}$.

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

*(ii)* $(\theta, \gamma) \mapsto \mathbb{E}^{F_*}[\phi^\star(a_1 + a_2 k(U, \theta, \gamma) + a_3' g(U, \theta, \gamma))]$ *is continuous 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* $(\theta, \gamma_0, P_0)$ *for each* $\theta \in \Theta_\delta(\gamma_0, P_0)$*;*

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

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

Parts (i) and (ii) of Assumption M are continuity conditions. If $k$ and $g$ consist of indicator functions, then these conditions hold provided the probabilities of the events under $F_*$ are continuous in $(\theta, \gamma)$. In models without $\gamma$, these conditions simply require continuity in $\theta$.

There are two parts to Assumption M(iii). The nonemptyness condition holds when the model is correctly specified under $F_*$ or, more generally, when there is at least one $F \in \mathcal{N}_\delta$ that satisfies (1) for some $\theta$. The second part is a constraint qualification. This condition requires that for each $\theta \in \Theta_\delta(\gamma_0, P_0)$, there is a distribution $F$ under which (1) holds at $(\theta, \gamma_0, P_0)$ that is "interior" to $\mathcal{N}_\infty$, in the sense that one can perturb the moments at $(\theta, \gamma_0, P_0)$ in all directions by perturbing $F$. Condition S' also requires that there is $F \in \mathcal{N}_\infty$ under which any inequality restrictions at $(\theta, \gamma_0, P_0)$ hold strictly. Note, however, that we do not require that this $F$ belongs to $\mathcal{N}_\delta$, only to $\mathcal{N}_\infty$. We therefore do not view this condition as overly restrictive. We also conjecture it could be relaxed using a notion similar to $S$-regularity from Section 2.5.

Assumption M(iv) is made for convenience and can be relaxed; this condition simply ensures that there do not exist values of $\theta$ at which $\Delta(\theta; \gamma_0, P_0) = \delta$ that are separated from $\Theta_\delta(\gamma_0, P_0)$. Assumption M(v) is standard and can be relaxed.

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

To derive the asymptotic distribution of the estimators, we assume $\gamma_0$ is known and suppress dependence of all quantities on $\gamma$ for the remainder of this section. This entails no loss of generality for models without $\gamma$, 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. The asymptotic distribution therefore reflects only sampling uncertainty from the estimated CCPs, which is the case for confidence sets reported when laws of motion are first estimated "offline". Extending our approach to accommodate sampling variation in $\hat{\gamma}$ in a tractable manner appears to require exploiting application-specific model structure, which we defer to future work.

Define

$$\underline{b}_\delta(P) = \inf_{\theta \in \Theta_\delta(P)} \underline{K}_\delta(\theta; P)\,, \qquad \overline{b}_\delta(P) = \sup_{\theta \in \Theta_\delta(P)} \overline{K}_\delta(\theta; P)\,. \tag{29}$$

In this notation, $\underline{\kappa}_\delta = \underline{b}_\delta(P_0)$ and $\overline{\kappa}_\delta = \overline{b}_\delta(P_0)$ (see Lemma E.3) and $\hat{\underline{\kappa}}_\delta = \underline{b}_\delta(\hat{P})$ and $\hat{\overline{\kappa}}_\delta = \overline{b}_\delta(\hat{P})$. We derive the asymptotic distribution of $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ by showing $\underline{b}_\delta$ and $\overline{b}_\delta$ are directionally differentiable and applying a suitable delta method. Say $f : \mathbb{R}^{d_1+d_2} \to \mathbb{R}$ 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

$$\lim_{n\to\infty} t_n^{-1} \left( f(P_0 + t_n h_n) - f(P_0) \right) = df_{P_0}[h]$$

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 $P_0$. We introduce some additional notation used to define the directional derivatives of $\underline{b}_\delta$ and $\overline{b}_\delta$. Let

$$\underline{\Xi}_\delta(\theta; P) = \text{argsup}_{\eta \ge 0, \zeta \in \mathbb{R}, \lambda \in \Lambda} - \mathbb{E}^{F_*}\left[ (\eta\phi)^\star(-k(U,\theta) - \zeta - \lambda' g(U,\theta)) \right] - \eta\delta - \zeta - \lambda'_{12} P\,,$$

where $(\eta\phi)^\star$ denotes the convex conjugate of $x \mapsto \eta \cdot \phi(x)$, and let $\overline{\Xi}_\delta(\theta; P)$ denote the analogous arginf for the minimization problem corresponding to the upper bound. Recall that $\underline{\lambda}_{12} = (\underline{\lambda}_1, \underline{\lambda}_2)$ collects the first $d_1 + d_2$ elements of $\underline{\lambda}$. Let

$$\underline{\Lambda}_\delta(\theta; P) = \{(\lambda_1, \lambda_2) : (\eta, \zeta, \lambda_1, \lambda_2, \lambda_3, \lambda_4) \in \underline{\Xi}_\delta(\theta; P)\}$$

denote the projection of $\underline{\Xi}_\delta(\theta; P)$ for $\underline{\lambda}_{12}$. We let $\overline{\Lambda}_\delta(\theta; P)$ denoting the analogous projection of $\overline{\Xi}_\delta(\theta; P)$. Finally, let

$$\underline{\Theta}_\delta(P_0) = \arg\min_{\theta\in\Theta} \underline{K}_\delta(\theta; P_0)\,, \quad \overline{\Theta}_\delta(P_0) = \arg\max_{\theta\in\Theta} \overline{K}_\delta(\theta; P_0)\,.$$

The sets $\underline{\Theta}_\delta(P_0)$ and $\overline{\Theta}_\delta(P_0)$ are nonempty and compact under Assumptions $\Phi$ and M.

The following regularity conditions are presented for the general case where $k$ depends on $u$. It may be possible to weaken some of these regularity conditions in the special case in which $k$ does not depend on $u$.

**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** *Suppose that Assumptions $\Phi$ and M hold. Then $\underline{b}_\delta$ and $\overline{b}_\delta$ are directionally differentiable at $P_0$, with*

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

*Moreover, if $\sqrt{n}(\hat{P} - P_0) \to_d Z \sim N(0, \Sigma)$ with $\Sigma$ finite, then*

$$\sqrt{n}\left(\begin{pmatrix} \hat{\overline{\kappa}}_\delta \\ \hat{\underline{\kappa}}_\delta \end{pmatrix} - \begin{pmatrix} \overline{\kappa}_\delta \\ \underline{\kappa}_\delta \end{pmatrix}\right) \to_d \begin{pmatrix} d\underline{b}_{\delta,P_0}[Z] \\ d\overline{b}_{\delta,P_0}[Z] \end{pmatrix}.$$

Theorem 6.2 derives the asymptotic distribution by extending a result of Shapiro (2008).[23] The asymptotic distribution will generally be non-Gaussian. If $\cup_{\theta\in\underline{\Theta}_\delta(P_0)}\underline{\Lambda}_\delta(\theta; P_0) = \{\underline{\lambda}_{12}\}$, then the asymptotic distribution of $\hat{\underline{\kappa}}_\delta$ simplifies to $\sqrt{n}(\hat{\underline{\kappa}}_\delta - \underline{\kappa}_\delta) \to_d N(0, \underline{\lambda}_{12}'\Sigma\underline{\lambda}_{12})$. An analogous simplification holds for $\hat{\overline{\kappa}}_\delta$ when $\cup_{\theta\in\overline{\Theta}_\delta(P_0)}\overline{\Lambda}_\delta(\theta; P_0)$ is a singleton.

[23] The extensions allow for non-compact domain for the inner problem, non-convexity/concavity in $\theta$, and discontinuity and unboundedness of the objective. In independent work, Galvao and Parker (2019) establish a similar result under the assumption that the objective function is bounded.

## 6.2 Inference Procedure 1: Bootstrap

Our first inference procedure specializes 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

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

with $\hat{\nu}$ a (possibly random) positive scalar tuning parameter for which $\hat{\nu} \to_p \nu > 0$. Any such $\hat{\nu}$ results in a confidence set with asymptotically correct coverage. We give some practical guidance for choosing $\hat{\nu}$ below.

Let $\hat{P}^*$ denote a bootstrapped version of $\hat{P}$. In practice any bootstrap can be used provided it satisfies mild consistency conditions. 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 $\Sigma$. Let

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

where the quantiles are computed by resampling $\hat{P}^*$. Lower, upper, and two-sided $100(1-\alpha)\%$ CSs for $\underline{\kappa}_\delta$, $\overline{\kappa}_\delta$, and $[\underline{\kappa}_\delta, \overline{\kappa}_\delta]$ are, respectively:

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

We require a slight strengthening of Assumption M(vii) to establish validity of the procedure. As before, regularity conditions are presented for the general case where $k$ depends on $u$. It may be possible to weaken these conditions when $k$ does not depend on $u$.

**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 $(\theta, P_0)$ for each $\theta \in \underline{\Theta}_\delta(P_0)$ and $\theta \in \overline{\Theta}_\delta(P_0)$, respectively.*

**Theorem 6.3** *Suppose that Assumptions $\Phi$ and M(i)–(vi),(vii') hold, $\sqrt{n}(\hat{P} - P_0) \to_d Z \sim N(0, \Sigma)$ with $\Sigma$ finite, and $\hat{P}^*$ satisfies Assumption 3 of Fang and Santos (2019). Then the distribution of $\widehat{d\underline{b}}_{\delta,P_0}[\sqrt{n}(\hat{P}^* - \hat{P})]$ and $\widehat{d\overline{b}}_{\delta,P_0}[\sqrt{n}(\hat{P}^* - \hat{P})]$ (conditional on the data) is consistent* *for the asymptotic distribution derived in Theorem 6.2. Moreover, if the CDFs of* $d\underline{b}_{\delta,P_0}[Z]$ *and* $d\bar{b}_{\delta,P_0}[Z]$ *are continuous and increasing at their* $\alpha/2$, $\alpha$, $1-\alpha$, *and* $1-\alpha/2$ *quantiles, then*

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

Any $\hat{\nu}$ that satisfies $\hat{\nu} \to_p \nu > 0$ results in asymptotically valid CSs. In view of the functional forms of $\widehat{d\bar{b}}_{\delta,P_0}[\,\cdot\,]$ and $\widehat{d\underline{b}}_{\delta,P_0}[\,\cdot\,]$, smaller $\hat{\nu}$ produce (weakly) wider CSs. In the CSW application, we set $\hat{\nu}$ equal to the minimum diagonal element of the covariance matrix of the moments evaluated at $(\hat{\theta}, \hat{\gamma}, \hat{P})$ under $F_*$, where $\hat{\theta}$ is computed under $F_*$. We chose this quantity as it is related to the convexity of the inner problem for small $\delta$. In practice, this resulted in $\hat{\nu}$ between 0.001 and 0.01. 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 $\hat{\nu}$. Setting $\hat{\nu} = 0$ 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 also valid, but may be conservative.

## 6.3 Inference Procedure 2: Projection

This second approach is computationally simple but possibly conservative.[24] 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 $P_0$:

$$\liminf_{n\to\infty} \Pr\left(P_{10} \leq \hat{P}_{1,U}^{1-\alpha}\,,\, \hat{P}_{2,L}^{1-\alpha} \leq P_{20} \leq \hat{P}_{2,U}^{1-\alpha}\right) \geq 1-\alpha\,, \tag{30}$$

where the inequalities should be understood to hold element-wise (we discuss how to construct a rectangular CS for $P_0$ below).

The idea behind this approach is to replace any moment conditions involving $P$ by inequalities constructed from the rectangular CS. Define the criterion functions

$$\underline{\hat{K}}_{\delta,1-\alpha}(\theta) = \begin{bmatrix} \underline{K}_{\delta,cs}(\theta;\hat{P}_{1-\alpha}) \\ +\infty \end{bmatrix}, \quad \hat{\overline{K}}_{\delta,1-\alpha}(\theta) = \begin{bmatrix} \overline{K}_{\delta,cs}(\theta;\hat{P}_{1-\alpha}) & \text{if } \Delta_{cs}(\theta;\hat{P}_{1-\alpha}) < \delta, \\ -\infty & \text{if } \Delta_{cs}(\theta;\hat{P}_{1-\alpha}) \geq \delta, \end{bmatrix}$$

where $\underline{K}_{\delta,cs}$, $\overline{K}_{\delta,cs}$, and $\Delta_{cs}$ are versions of (13), (14), and (16) formed using the inequalities

$$\mathbb{E}^F[g_1(U,\theta)] \leq \hat{P}_{1,U}^{1-\alpha}\,, \quad \mathbb{E}^F[g_2(U,\theta)] \leq \hat{P}_{2,U}^{1-\alpha}\,, \quad \mathbb{E}^F[-g_2(U,\theta)] \leq -\hat{P}_{2,L}^{1-\alpha}\,, \tag{31}$$

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

as well as (1c) and (1d). In these criterions, $\Lambda$ is replaced by $\Lambda_{cs} = \mathbb{R}_+^{d_1+2d_2+d_3} \times \mathbb{R}^{d_4}$, $g$ 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 $\lambda_{12}$ denotes the first $d_1 + 2d_2$ elements of $\lambda$.

Critical values are computed by optimizing the criterions $\hat{\underline{K}}_{\delta,1-\alpha}$ and $\hat{\overline{K}}_{\delta,1-\alpha}$ with respect to $\theta$:

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

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

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

**Theorem 6.4** *Suppose that Assumptions $\Phi$ and M(i),(iii)–(v) hold and $\hat{P}_{1-\alpha}$ satisfies (30). Then*

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

To construct a rectangular CS for $P_0$ satisfying (30), suppose $\sqrt{n}(\hat{P} - P_0) \to_d N(0,\Sigma)$ and we have a consistent estimator $\hat{\Sigma}$ of $\Sigma$. Let $\hat{\sigma}$ denote the vector formed by taking the square root of each diagonal entry of $\hat{\Sigma}$. 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

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

If $d_2 = 0$, then $\hat{c}_{1-\alpha,1}$ is the $(1-\alpha)$-quantile of $\max_{1\leq i\leq d_1} Z_i/\hat{\sigma}_i$; similarly, if $d_1 = 0$, then $\hat{c}_{2,1-\alpha}$ is the $(1-\alpha)$-quantile of $\max_{1\leq i\leq d_2} |Z_i/\hat{\sigma}_i|$.

# 7 Conclusion

This paper introduced a framework for performing a global sensitivity analysis of counterfactuals to parametric assumptions about the distribution of latent variables in structural models. In particular, we derived bounds on the set of counterfactuals obtained as the distribution of latent variables spans nonparametric neighborhoods of the researcher's chosen parametric specification while other "structural" model features are maintained. We illustrated our procedure with empirical applications to matching models and dynamic discrete choice.

# References


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.

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

Berry, S. T. and P. A. Haile (2010). Nonparametric identification of multinomial choice demand models with heterogeneous consumers. Cowles foundation discussion paper no. 1718.

Berry, S. T. and P. A. Haile (2014). Identification in differentiated products markets using market level data. *Econometrica 82*(5), 1749–1797.

Bonhomme, S. and M. Weidner (2021). Minimizing sensitivity to model misspecification. *Quantitative Economics*, forthcoming.

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., T. M. Christensen, and E. Tamer (2018). Monte carlo confidence sets for identified sets. *Econometrica 86*(6), 1965–2018.

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.

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. (1975). $I$-divergence geometry of probability distributions and minimization problems. *The Annals of Probability 3*(1), 146–158.

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

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems 26*, 2292–2300.

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

Duchi, J. C. and H. Namkoong (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics 49*(3), 1378–1406.

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 (2021). Cupid's invisible hand: Social surplus and identification in matching models. *The Review of Economic Studies*, forthcoming.

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.

Gualdani, C. and S. Sinha (2020). Partial identification in matching models for the marriage market. arxiv:1902.05610 [econ.em].

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.

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 (2021). Identification of counterfactuals in dynamic discrete choice models. *Quantitative Economics 12*(2), 351–403.

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.

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.

Milgrom, P. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *Econometrica 70*(2), 583–601.

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.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

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.

Shapiro, A. (2017). Distributionally robust stochastic programming. *SIAM Journal on Optimization 27*(4), 2258–2275.

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.

Tamer, E. (2015). Sensitivity analysis in some econometric models. Cowles Lecture, Econometric Society World Congress, August 17-21, 2015, Montréal.

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

This supplement presents extensions of our methodology in Appendix A, additional results on nonparametric bounds on counterfactuals in Appendix B, connections with local approaches to sensitivity analysis in Appendix C, additional details on the empirical applications in Appendix D, and proofs of results from the main text in Appendix E.

# A Extensions

This appendix presents three extensions of our methodology. Proofs of all results in this appendix are deferred to Appendix G.6.

## A.1 Group Invariance

In certain settings it can be attractive to impose shape restrictions on $F$ such as symmetry, exchangeability, or, more generally, invariance to a finite group of transforms. For instance, imposing exchangeability of $F$ in discrete choice modeling ensures that alternatives' choice probabilities depend on their deterministic components of utility but not their labeling. These shape restrictions can be easily imposed whenever $F_*$ is invariant.

Formally, let $J$ denote the number of elements of $U$ and let $\Pi$ be a finite commutative group of transforms on $\mathbb{R}^J$—see, e.g., Section 1.4 of Lehmann and Casella (1998). We say that a distribution $F$ of $U$ is $\Pi$-*invariant* if $\varpi U \sim F$ for all $\varpi \in \Pi$.

**Example A.1 (Symmetry)** Central symmetry corresponds to $\Pi = \{I, -I\}$ for $I$ the identity matrix. Sign symmetry corresponds to taking $\Pi$ to be the collection of all $2^J$ diagonal matrices with $\pm 1$ in each diagonal entry. □

**Example A.2 (Exchangeability)** Let $\Pi_J$ denote the group of all $J!$ permutation matrices of dimension $J$. Full exchangeability (permutation invariance) corresponds to $\Pi = \Pi_J$. Cyclic exchangeability (rotation invariance) corresponds to $\Pi = \Pi_J^c$ where $\Pi_J^c$ is the collection of all $J$ cyclic permutation matrices of dimension $J$ ($\Pi_J^c = \Pi_J$ when $J = 2$ and is a strict subset otherwise). When $J \geq 3$, dihedral exchangeability (rotation and reflection invariance) corresponds to taking $\Pi$ to be the set of all $2J$ permutation matrices representing rotations and reflections of $\{1, \ldots, J\}$. These types of exchangeability ensure the elements of $U$ are identically distributed, but they have different implications for the joint distribution of the 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, but is independent of $(i, j)$ under full exchangeability. $\square$

Let $\mathcal{N}_\delta^\Pi = \{F \in \mathcal{N}_\delta : F \text{ is } \Pi\text{-invariant}\}$. We are interested in

$$\underline{\kappa}_\delta^\Pi := \inf_{\theta \in \Theta, F \in \mathcal{N}_\delta^\Pi} \mathbb{E}^F[k(U, \theta, \gamma_0)] \quad \text{subject to (1)}, \tag{32}$$

and $\overline{\kappa}_\delta^\Pi$ defined as the analogous supremum. One may write $\underline{\kappa}_\delta^\Pi$ and $\overline{\kappa}_\delta^\Pi$ as the value of two optimization problems in which criterion functions $\underline{K}_\delta^\Pi(\theta; \gamma_0, P_0)$ and $\overline{K}_\delta^\Pi(\theta; \gamma_0, P_0)$ are optimized with respect to $\theta$. For a generic $(\theta, \gamma, P)$, define

$$\underline{K}_\delta^\Pi(\theta; \gamma, P) = \inf_{F \in \mathcal{N}_\delta^\Pi} \mathbb{E}^F[k(U, \theta, \gamma)] \quad \text{subject to (1) holding at } (\theta, \gamma, P)\,, \tag{33}$$

and define $\overline{K}_\delta^\Pi(\theta; \gamma, P)$ as the analogous supremum. These criterions have dual representations as finite-dimensional convex programs when $F_*$ is $\Pi$-invariant. Define

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

where $|\Pi|$ denotes the cardinality of $\Pi$, and let $g^\Pi = (g_1^\Pi, g_2^\Pi, g_3^\Pi, g_4^\Pi)$.

**Proposition A.1** *Suppose that Assumption $\Phi$ holds and $F_*$ is $\Pi$-invariant. Then*

$$\underline{K}_\delta^\Pi(\theta; \gamma, P) = \sup_{\eta > 0, \zeta \in \mathbb{R}, \lambda \in \Lambda} -\eta \mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k^\Pi(U,\theta,\gamma) + \zeta + \lambda' g^\Pi(U,\theta,\gamma)}{-\eta}\right)\right] - \eta\delta - \zeta - \lambda_{12}' P\,, \tag{34}$$

$$\overline{K}_\delta^\Pi(\theta; \gamma, P) = \inf_{\eta > 0, \zeta \in \mathbb{R}, \lambda \in \Lambda} \eta \mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k^\Pi(U,\theta,\gamma) - \zeta - \lambda' g^\Pi(U,\theta,\gamma)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda_{12}' P\,. \tag{35}$$

*Moreover, the value of problem (34) is $+\infty$ (equivalently, the value of problem (35) is $-\infty$) if and only if there is no distribution in $\mathcal{N}_\delta^\Pi$ under which (1) holds at $(\theta, \gamma, P)$.*

**Remark A.1** *If $F$ is $\Pi$-invariant and satisfies (1), then it must also satisfy (1) under all $|\Pi|$ transformations of the elements of $U$. Therefore, in effect there are a total of $|\Pi| \times d$ moment conditions imposed in the inner optimization, namely*

$$\begin{aligned} &\mathbb{E}^F[g_1(\varpi U, \theta, \gamma_0)] \leq P_{10}, \quad \mathbb{E}^F[g_2(\varpi U, \theta, \gamma_0)] = P_{20}, \\ &\mathbb{E}^F[g_3(\varpi U, \theta, \gamma_0)] \leq 0, \qquad \mathbb{E}^F[g_4(\varpi U, \theta, \gamma_0)] = 0, \end{aligned} \qquad \textit{for all } \varpi \in \Pi\,. \tag{36}$$

*In principle one could form a criterion by including all $|\Pi| \times d$ moments. By $\Pi$-invariance of $F_*$ and convexity of the objective, the multipliers on the moments $g(\varpi U, \theta, \gamma)$ will be identical across all $\varpi \in \Pi$. It therefore suffices to form the criterion using only the $d$ averaged moments $g^\Pi$ rather than the full set of $|\Pi| \times d$ moments, thereby reducing the dimension of the inner optimization by a factor of $|\Pi|$.*

**Remark A.2** *When Monte Carlo integration is used to compute expectations, taking a sample from $F_*$ and then concatenating the sample across each of its $|\Pi|$ transformations ensures the empirical distribution of the random draws is $\Pi$-invariant.*

## A.2 Conditional Moment Models

Consider the conditional moment model

$$\begin{array}{ll} \mathbb{E}^F[g_1(U,X,\theta,\gamma_0)|X=x] \leq P_{10,x}, & \mathbb{E}^F[g_2(U,X,\theta,\gamma_0)|X=x] = P_{20,x}, \\ \mathbb{E}^F[g_3(U,X,\theta,\gamma_0)|X=x] \leq 0, & \mathbb{E}^F[g_4(U,X,\theta,\gamma_0)|X=x] = 0, \end{array} \quad \text{for all } x \in \mathcal{X} \tag{37}$$

where $\mathcal{X}$ is a finite set, and a counterfactual[25]

$$\kappa = \sum_{x \in \mathcal{X}} \mathbb{E}^F[k(U,X,\theta,\gamma_0)|X=x]\,. \tag{38}$$

Suppose the researcher assumes $U|X = x \sim F_*$ for each $x$. We wish to relax this assumption and allow each conditional distribution of $U$ given $X = x$, say $F_x$, to vary in a neighborhood $\mathcal{N}_{\delta_x}$ of $F_*$. In doing so, we are allowing the conditional distributions $F_x$ to vary with $x$, and therefore relaxing independence of $U$ and $X$.[26]

We assume each $\mathcal{N}_\delta$ is defined by the same $\phi$ to simplify the exposition, but we allow the neighborhood size to vary with $x$. Let $\boldsymbol{\delta} = (\delta_x)_{x\in\mathcal{X}}$. We are interested in

$$\underline{\kappa}_{\boldsymbol{\delta}} := \inf_{\theta\in\Theta,(F_x\in\mathcal{N}_{\delta_x})_{x\in\mathcal{X}}} \sum_{x\in\mathcal{X}} \mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{subject to (37)}, \tag{39}$$

and $\overline{\kappa}_{\boldsymbol{\delta}}$ defined as the analogous supremum. One may write $\underline{\kappa}_{\boldsymbol{\delta}}$ and $\overline{\kappa}_{\boldsymbol{\delta}}$ as the value of two optimization problems where $\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma_0,P_0)$ and $\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma_0,P_0)$ are optimized with respect to $\theta$.

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

[26] 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, as in Examples 2.1–2.3.

Let $P = (P_x)_{x \in \mathcal{X}}$ where $P_x = (P_{1,x}, P_{2,x})$ is partitioned conformably with $g_1$ and $g_2$. For a generic $(\theta, \gamma, P)$, define

$$\underline{K}_{\boldsymbol{\delta}}(\theta; \gamma, P) = \inf_{(F_x \in \mathcal{N}_{\delta_x})_{x \in \mathcal{X}}} \sum_{x \in \mathcal{X}} \mathbb{E}^{F_x}[k(U, x, \theta, \gamma_0)] \quad \text{subject to (37) holding at } (\theta, \gamma, P),$$

and define $\overline{K}_{\boldsymbol{\delta}}(\theta; \gamma, P)$ as the analogous supremum. These criterion functions have dual forms analogous to Proposition 2.1. Let $g(\cdot, x, \theta, \gamma) = (g_1(\cdot, x, \theta, \gamma), \ldots, g_4(\cdot, x, \theta, \gamma))$. Recall $d = \sum_{i=1}^{4} d_i$ where $d_i$ is the dimension of $g_i$, and $\Lambda = \mathbb{R}_+^{d_1} \times \mathbb{R}^{d_2} \times \mathbb{R}_+^{d_3} \times \mathbb{R}^{d_4}$. Let $\lambda_{12,x}$ denote the first $d_1 + d_2$ elements of $\lambda_x \in \Lambda$.

**Assumption $\Phi$-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* $(\theta, \gamma, x) \in \Theta \times \Gamma \times \mathcal{X}$.

**Proposition A.2** *Suppose that Assumption $\Phi$-conditional holds. Then*

$$\begin{aligned} &\underline{K}_{\boldsymbol{\delta}}(\theta; \gamma, P) && (40) \\ &= \sup_{(\eta_x > 0, \zeta_x \in \mathbb{R}, \lambda_x \in \Lambda)_{x \in \mathcal{X}}} \sum_{x \in \mathcal{X}} \left( -\eta_x \mathbb{E}^{F_*} \left[ \phi^\star \left( \frac{k(U,x,\theta,\gamma) + \zeta_x + \lambda_x' g(U,x,\theta,\gamma)}{-\eta_x} \right) \right] - \eta_x \delta_x - \zeta_x - \lambda_{12,x}' P_x \right), \\ &\overline{K}_{\boldsymbol{\delta}}(\theta; \gamma, P) && (41) \\ &= \inf_{(\eta_x > 0, \zeta_x \in \mathbb{R}, \lambda_x \in \Lambda)_{x \in \mathcal{X}}} \sum_{x \in \mathcal{X}} \left( \eta_x \mathbb{E}^{F_*} \left[ \phi^\star \left( \frac{k(U,x,\theta,\gamma) - \zeta_x - \lambda_x' g(U,x,\theta,\gamma)}{\eta_x} \right) \right] + \eta_x \delta_x + \zeta_x + \lambda_{12,x}' P_x \right). \end{aligned}$$

*Moreover, the value of (40) is $+\infty$ (equivalently, the value of (41) is $-\infty$) if and only if for some $x \in \mathcal{X}$ there is no distribution in $\mathcal{N}_{\delta_x}$ under which (37) holds at $(\theta, \gamma, P)$.*

As before, 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 optimizing sample criterions with respect to $\theta$. Let $\hat{P} = (\hat{P}_x)_{x \in \mathcal{X}}$. The sample criterions are

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

where $\overline{K}_{\boldsymbol{\delta}}(\theta; \hat{\gamma}, \hat{P})$ and $\overline{K}_{\boldsymbol{\delta}}(\theta; \hat{\gamma}, \hat{P})$ denote the programs in Proposition A.2 evaluated at $(\hat{\gamma}, \hat{P})$, and

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

## A.3 Non-separable Models

Consider the model

$$\begin{aligned} &\mathbb{E}^H[\tilde{g}_1(U,X,\theta,\tilde{\gamma}_0)] \le P_{10}, \quad \mathbb{E}^H[\tilde{g}_2(U,X,\theta,\tilde{\gamma}_0)] = P_{20}, \\ &\mathbb{E}^H[\tilde{g}_3(U,X,\theta,\tilde{\gamma}_0)] \le 0, \qquad \mathbb{E}^H[\tilde{g}_4(U,X,\theta,\tilde{\gamma}_0)] = 0, \end{aligned} \tag{42}$$

and counterfactual

$$\kappa = \mathbb{E}^H[\tilde{k}(U,X,\theta,\tilde{\gamma}_0)]\,, \tag{43}$$

where the expectation is with respect to the distribution $H$ of $(U,X)$ and $X$ takes values in a finite set $\mathcal{X}$. Suppose the researcher assumes $U|X=x \sim F_*$ for each $x$. We wish to relax this assumption and allow the conditional distribution of $U$ given $X=x$, say $F_x$, to vary in a neighborhood $\mathcal{N}_{\delta_x}$ of $F_*$.

Write $H(u,x) = q_{0,x} \cdot F_x(u)$ where $q_{0,x} = \Pr(X=x)$. The vector $q_0 = (q_{0,x})_{x\in\mathcal{X}}$ can be consistently estimated from data on $X$. Let $\gamma_0 = (\tilde{\gamma}_0, q_0)$. Define $g_1(U,x,\theta,\gamma_0) = q_{0,x} \cdot \tilde{g}_1(U,x,\theta,\tilde{\gamma}_0)$ and similarly for $g_2$, $g_3$, $g_4$, and $k$. The model (42) and counterfactual (43) can then be written as

$$\begin{aligned} &\sum_x \mathbb{E}^{F_x}[g_1(U,x,\theta,\gamma_0)] \le P_{10}, \quad \sum_x \mathbb{E}^{F_x}[g_2(U,x,\theta,\gamma_0)] = P_{20}, \\ &\sum_x \mathbb{E}^{F_x}[g_3(U,x,\theta,\gamma_0)] \le 0, \qquad \sum_x \mathbb{E}^{F_x}[g_4(U,x,\theta,\gamma_0)] = 0, \end{aligned} \tag{44}$$

and $\kappa = \sum_x \mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)]$. We again assume each $\mathcal{N}_\delta$ is defined by the same $\phi$ function, but allow the neighborhood size to vary with $x$. Let $\boldsymbol{\delta} = (\delta_x)_{x\in\mathcal{X}}$. We are interested in

$$\underline{\kappa}_{\boldsymbol{\delta}} := \inf_{\theta\in\Theta,(F_x\in\mathcal{N}_{\delta_x})_{x\in\mathcal{X}}} \sum_x \mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{subject to (44)},$$

and $\overline{\kappa}_{\boldsymbol{\delta}}$ defined as the analogous supremum. One may write $\underline{\kappa}_{\boldsymbol{\delta}}$ and $\overline{\kappa}_{\boldsymbol{\delta}}$ as the value of two optimization problems where criterion functions $\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma_0,P_0)$ and $\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma_0,P_0)$ are optimized with respect to $\theta$. For a generic $(\theta,\gamma,P)$, define

$$\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P) = \inf_{(F_x\in\mathcal{N}_{\delta_x})_{x\in\mathcal{X}}} \sum_x \mathbb{E}^{F_x}[k(U,x,\theta,\gamma_0)] \quad \text{s.t. (44) holding at } (\theta,\gamma,P)\,,$$

and define $\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P)$ as the analogous supremum. These criterion functions have dual forms analogous to Proposition 2.1. Let $g(\cdot,x,\theta,\gamma) = (g_1(\cdot,x,\theta,\gamma),\ldots,g_4(\cdot,x,\theta,\gamma))$. The remaining notation the same as Proposition 2.1.

**Proposition A.3** *Suppose that Assumption* $\Phi$*-conditional holds. Then*

$$\underline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P) \tag{45}$$
$$= \sup_{(\eta_x>0,\zeta_x\in\mathbb{R})_{x\in\mathcal{X}},\lambda\in\Lambda} \sum_x \left(-\eta_x \mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,x,\theta,\gamma)+\zeta_x+\lambda' g(U,x,\theta,\gamma)}{-\eta_x}\right)\right] - \eta_x\delta_x - \zeta_x - \lambda_{12}' P\right),$$
$$\overline{K}_{\boldsymbol{\delta}}(\theta;\gamma,P) \tag{46}$$
$$= \inf_{(\eta_x>0,\zeta_x\in\mathbb{R})_{x\in\mathcal{X}},\lambda\in\Lambda} \sum_x \left(\eta_x \mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,x,\theta,\gamma)-\zeta_x-\lambda' g(U,x,\theta,\gamma)}{\eta_x}\right)\right] + \eta_x\delta_x + \zeta_x + \lambda_{12}' P\right).$$

*Moreover, the value of (45) is* $+\infty$ *(equivalently, the value of (46) is* $-\infty$*) if and only if there is no* $H(u,x) = q_{0,x}\cdot F_x(u)$ *with* $F_x \in \mathcal{N}_{\delta_x}$ *under which (42) holds at* $(\theta,\gamma,P)$*.*

As before, 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 optimizing sample criterion functions with respect to $\theta$. The sample criterion functions are

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

where $\overline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P)$ and $\overline{K}_{\boldsymbol{\delta}}(\theta;\hat\gamma,\hat P)$ denote the programs in Proposition A.3 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

$$\Delta_{nonsep}(\theta;\gamma,P)$$
$$= \sup_{\substack{(\eta_x\ge 0,\zeta_x\in\mathbb{R})_{x\in\mathcal{X}},\lambda\in\Lambda \\ \sum_{x\in\mathcal{X}}\eta_x\le 1}} \left(-\sum_{x\in\mathcal{X}} \mathbb{E}^{F_*}\left[(\eta_x\phi)^\star(-\zeta_x - \lambda_x' g(U,x,\theta,\gamma))\right] - \eta_x\delta_x - \zeta_x\right) - \lambda_{12}' P\,.$$

By similar arguments to Appendix G.3, $\Delta_{nonsep}(\theta;\gamma,P)$ may be shown to be the dual of

$$\inf_{t\in\mathbb{R},(F_x)_{x\in\mathcal{X}}} t \quad \text{s.t.} \quad D_\phi(F_x\|F_*) \le \delta_x + t \text{ for each } x\in\mathcal{X} \text{ and (44) holding at } (\theta,\gamma,P).$$

Therefore, if there exists $F_x$ with $D_\phi(F_x\|F_*) < \delta_x$ for each $x$ such that (44) holds at $(\theta,\gamma,P)$, then $\Delta_{nonsep}(\theta;\gamma,P) < 0$.

# B Additional Results on Nonparametric Bounds

This appendix presents further details to supplement Section 2.5. Proofs of all results in this appendix are presented in Appendix G.7. Our first result concerns the behavior of $\underline{\kappa}_\delta$ and $\overline{\kappa}_\delta$

as the neighborhood size $\delta$ becomes large. Recall $\mathcal{N}_\infty = \{F : D_\phi(F\|F_*) < \infty\}$. Let

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

**Lemma B.1** *Suppose that Assumption Φ holds. Then*

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

Next, we characterize bounds on $\mathcal{K}_\infty$ using profiled optimization problems and derive their dual forms. Define

$$\underline{K}_\infty(\theta;\gamma_0,P_0) = \inf_{F\in\mathcal{N}_\infty} \mathbb{E}^F[k(U,\theta,\gamma_0)] \quad \text{subject to (1) holding at } (\theta,F)\,, \tag{47}$$

and let $\overline{K}_\infty(\theta;\gamma_0,P_0)$ denote the analogous supremum. By definition, we have

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

Let $F_*$-ess inf and $F_*$-ess sup denote the $F_*$-essential infimum and supremum, respectively

**Lemma B.2** *Suppose that Assumption Φ holds and Condition S holds at $(\theta,\gamma,P)$. Then*

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

We now derive analogous dual representations for the criterion functions $\underline{K}_{np}$ and $\overline{K}_{np}$ from Section 2.5 (see display (18)). We require a slightly different constraint qualification:

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

If $F_*$ and $\mu$ are mutually absolutely continuous, then Condition $\mathrm{S}_{np}$ is equivalent to Condition S from Section 2.5 (see Lemma E.1).

**Lemma B.3** *Suppose that Condition $S_{np}$ holds at $(\theta,\gamma,P)$ and $k$ is $\mu$-essentially bounded. Then*

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

# C Local Sensitivity

In this appendix, we first introduce a measure of local sensitivity of the counterfactual with respect to $F$. We then contrast our approach with recent work on local sensitivity.

## C.1 Measure of Local Sensitivity

Our *measure of local sensitivity* of the counterfactual $\kappa$ 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

$$\underline{\kappa}_\delta = \kappa_* - \sqrt{\delta s} + o(\sqrt{\delta})\,, \quad \overline{\kappa}_\delta = \kappa_* + \sqrt{\delta s} + o(\sqrt{\delta})\,, \quad \text{as } \delta \downarrow 0,$$

where $\kappa_* = \mathbb{E}^{F_*}[k(U, \theta_*, \gamma_0)]$ and $\theta_*$ solves (1) under $F_*$.

To draw connections with the local sensitivity literature, we restrict attention to moment equality models and impose (standard) regularity conditions. These conditions allow us to characterize $s$ very tractably via an influence function representation, which leads to a simple estimator $\hat{s}$ of $s$. Assume that under $F_*$ the moment conditions (1b) and (1d) point identify a structural parameter $\theta_* \in \mathrm{int}(\Theta)$, where we again assume $\Theta$ is compact. With some abuse of notation, let

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

$g_*(u) = g(u, \theta_*, \gamma_0, P_{20})$, 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 $\theta$ at $\theta_*$, $G = \frac{\partial}{\partial \theta'} \mathbb{E}^{F_*}[g(U, \theta, \gamma_0, P_{20})]\big|_{\theta=\theta_*}$ 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 $\theta$ at $\theta_*$.

Define the *influence function* of the counterfactual $\kappa$ with respect to $F$ at $F_*$ as

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

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 $J = \frac{\partial}{\partial \theta} \mathbb{E}^{F_*}[k(U, \theta, \gamma_0)]\big|_{\theta=\theta_*}$. The following theorem relates $s$ and $\iota$. We restrict attention to neighborhoods characterized by $\chi^2$ divergence. Other $\phi$-divergences are locally equivalent to

$\chi^2$ divergence, so this restriction entails no great loss of generality.[27]

**Theorem C.1** *Suppose that the above GMM-type regularity conditions hold and neighborhoods are defined using $\chi^2$ divergence. Then $s = 2\mathbb{E}^{F_*}[\iota(U)^2]$.*

The proof of Theorem C.1 is presented in Appendix G.8. In addition to reporting an estimated counterfactual $\hat{\kappa} = \mathbb{E}^{F*}[k(U, \hat{\theta}, \hat{\gamma})]$, researchers could also report an estimate of local sensitivity 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 $\hat{G} = \frac{\partial}{\partial\theta'}\mathbb{E}^{F_*}[g(U, \theta, \hat{\gamma}, \hat{P}_2)]|_{\theta=\hat{\theta}}$ and $\hat{J} = \frac{\partial}{\partial\theta}\mathbb{E}^{F_*}[k(U, \theta, \hat{\gamma})]|_{\theta=\hat{\theta}}$. Lemma G.10 in Appendix G.8 shows $\hat{s}$ is consistent. Bounds on counterfactuals as $F$ varies over small neighborhoods of $F_*$ can then be estimated using $\hat{\kappa} \pm \sqrt{\delta\hat{s}}$.

## C.2 Comparison with Other Approaches

We now compare our approach with Andrews et al. (2017, 2020), henceforth AGS, and Bonhomme and Weidner (2021), henceforth BW. To simplify the comparison, we consider models characterized by moments of the form (1b) with $d_2 \geq d_\theta$ and in which there is no $\gamma$.

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

$$\mathbb{E}^{F_*}[g_2(U, \theta_*)] = P_{20} + n^{-1/2}c\,. \tag{48}$$

Suppose a researcher has a first-stage estimator $\hat{P}_2$, computes an estimator $\hat{\theta}$ by minimizing

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

then estimates the counterfactual using $\hat{\kappa} = \mathbb{E}^{F_*}[k(U, \hat{\theta})]$. AGS's measure of *sensitivity* of $\hat{\kappa}$ to $\hat{P}_2$ is $J'(G'WG)^{-1}G'W$, where $W$ is the probability limit of $\hat{W}$. The first-order asymptotic bias of $\hat{\kappa}$ due to local misspecification is therefore $J'(G'WG)^{-1}G'Wc$. AGS's measure of

[27] 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 $\phi$ divergences.

*informativeness* of $\hat{P}_2$ for $\hat{\kappa}$ is 1, meaning that all sampling variation in $\hat{\kappa}$ is explained by sampling variation in $\hat{P}_2$. Our measure $s$ instead characterizes "specification variation" in $\kappa$ as the researcher varies $F$ subject to the moment condition (1b).

BW consider estimation of a target parameter ($\kappa$ in our context) using a reference model $\mathcal{M}_R = \{(\theta, F) \in \Theta \times \{F_*\}\}$ when the true $(\theta_0, F_0)$ possibly belongs to a larger model $\mathcal{M}_L = \{(\theta, F) \in \Theta \times \mathcal{N}_\delta\}$ with $\delta \downarrow 0$ as the sample size $n$ increases so that $n\delta \to \tau \geq 0$. BW seek estimators of $\kappa$ 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 $\theta_*$ 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 $\mathcal{M}_R$, so 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 $k + a'g_2$.

# D Additional Details for the Empirical Applications

## D.1 Marital College Premium

**Bootstrap Details.** CSs reported Section 5.1 with $\delta > 0$ are computed using the bootstrap procedure from Section 6.2. To implement the bootstrap, we take 1,000 independent draws of $\hat{P}_2^* \sim N(\hat{P}_2, \hat{\Sigma})$ where $\hat{\Sigma}$ is CSW's estimate of the covariance matrix of the sampling distribution of $\hat{P}_2$. We compute $\hat{P}_2$ and $\hat{\Sigma}$ based on CSW's replication files.

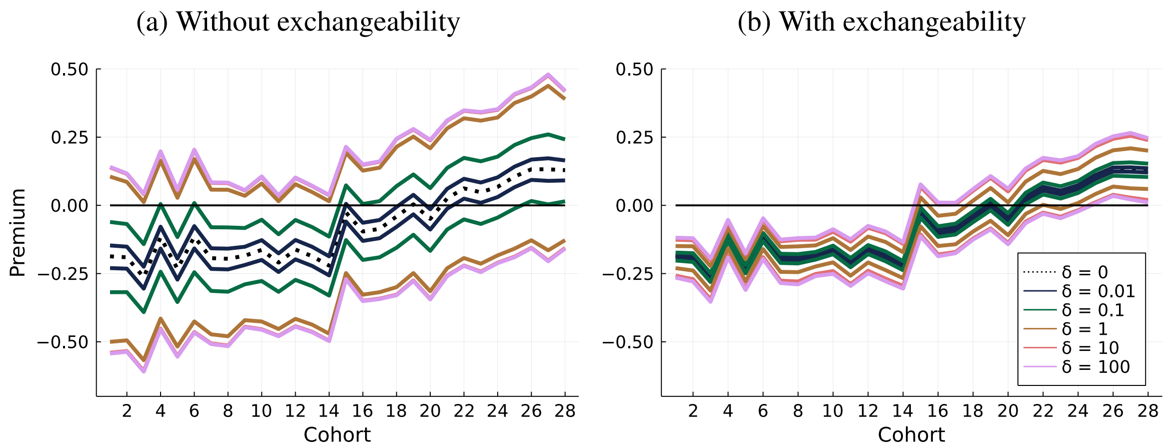


Figure 5: Fixed-$\theta$ bounds on the "some college" to "college graduate" premium when structural parameters are held fixed at CSW's estimates.

**Fixed-$\theta$ Bounds.** Figure 5a plots lower and upper bounds on the "some college" to "college graduate" premium across cohorts when $\theta$ is fixed at CSW's estimates (computed under $F_*$) but $F$ is allowed to vary. These bounds for large $\delta$ contain zero across each cohort, and are approximately the same width as the bounds with $\delta = 0.01$ reported in Figure 1a where both $\theta$ and $F$ are allowed to vary. Imposing exchangeability (Figure 5b) is seen to tighten the bounds substantially, producing bounds that span negative values only for early cohorts and positive values only in the latest few cohorts.

**Projection CSs.** Figure 6 reports projection CSs computed using the procedure in Section 6.3. We formed 95% rectangular CSs for each cohort's $P_{20}$ as described in Section 6.3 using CSW's estimates for $\hat{P}_2$ and their asymptotic variance estimates for $\hat{\Sigma}$. These CSs are significantly wider than the bootstrap CSs reported in Figure 1. Some conservativeness is to be expected, as these CSs project a 95% $P_{20}$ down to one dimension. The relative inefficiency is especially pronounced for the earlier cohorts, for which the sample size is smaller. Note also from Figure 6b that the projection CSs with $\delta = 0.01$ span zero across each cohort, whereas in Figure 1b the bootstrap CSs with $\delta = 0.01$ contain negative values only in some early cohorts and positive values only in later cohorts.

**Computation Times.** Table 5 reports times for solving the inner problem for maximizing the premium in cohort 1. This optimization problem defines the criterion function $\overline{K}_\delta(\theta; \hat{P})$. As times vary with $\theta$, we report times at CSW's estimates. Times increase somewhat with $\delta$, but are all under $0.6$ seconds. The outer optimization times varied with cohort, $\delta$, and implementation but were typically solved in at most a few minutes (often under 90 seconds).

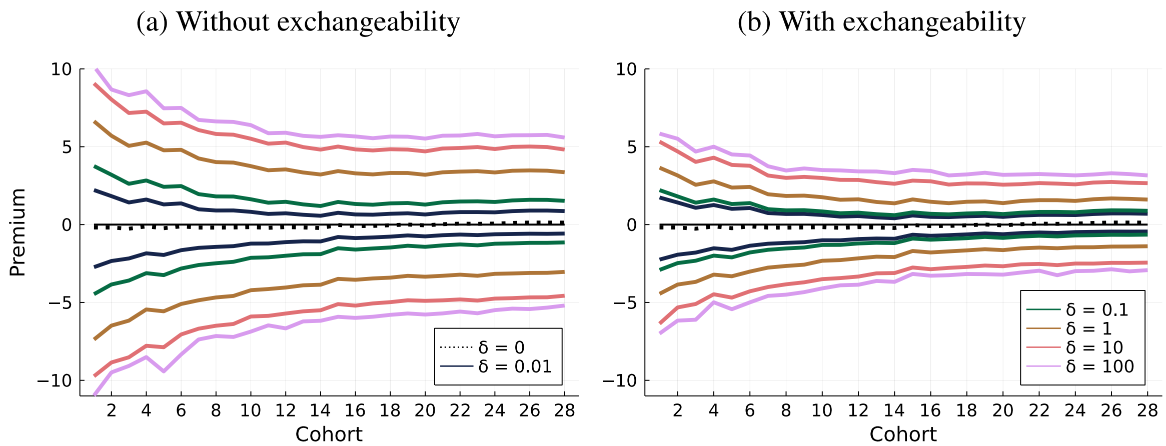


Figure 6: Projection 95% CSs for bounds on the "some college" to "college graduate" premium across cohorts.

Table 5: Computation times for the inner problem in the matching application

| Implementation | $\delta$ | | | | |
|---|---|---|---|---|---|
| | 0.01 | 0.1 | 1 | 10 | 100 |
| Without exchangeability | 0.074 | 0.056 | 0.076 | 0.579 | 0.188 |
| With exchangeability | 0.146 | 0.184 | 0.350 | 0.311 | 0.488 |

*Note:* Times (in seconds) for solving the inner optimization problem for maximizing the premium in cohort 1 at CSW's parameter estimate $\theta$. We use 50,000 Monte Carlo draws without exchangeability and 120,000 draws with exchangeability. All computations are performed in Julia version 1.6.4 and KNITRO 12.4.0 on a 2.7GHz MacBook Pro with 16GB memory.

**Sensitivity to $\phi$.** Using $\chi^2$ and $L^4$ divergences produced near identical bounds for $\delta = 0.01$ and $0.1$. The $\chi^2$ bounds with $\delta = 1$ and $10$ were at most 10% narrower than the hybrid bounds. The $L^4$ bounds were 60%-70% of the width of the hybrid bounds for $\delta = 1$, $10$, and $100$ across cohorts ($L^4$ divergence is stronger than $\chi^2$ and hybrid divergence). The shapes of the sets 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 $\phi$.

## D.2 Welfare Analysis in a Rust Model

**Bootstrap Details.** CSs reported Section 5.2 with $\delta > 0$ are computed using the bootstrap procedure from Section 6.2. To implement the bootstrap, we take 1,000 independent draws of $\hat{\theta}^*_\pi \sim N(\hat{\theta}_\pi, \hat{\Sigma})$ where $\hat{\theta}_\pi$ is the MLE of $(RC, MC)$ under the i.i.d. Gumbel assumption and $\hat{\Sigma}$ is an estimate of the inverse information matrix. We then set $\hat{P}^*_2$ to be the model-implied CCPs at $\hat{\theta}^*_\pi$ under the i.i.d. Gumbel assumption. As the function $k$ depends only implicitly on

Table 6: Computation times for the inner problem in the DDC application

| | $\delta$ | | | | |
|---|---|---|---|---|---|
| | 0.01 | 0.1 | 1 | 10 | 100 |
| Lower bound | 0.124 | 0.144 | 0.164 | 0.285 | 0.265 |
| Upper bound | 0.101 | 0.119 | 0.142 | 0.266 | 1.039 |

*Note:* Times (in seconds) for solving the inner optimization problem at the parameter values at which $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained. All computations are performed in Julia version 1.6.4 and KNITRO 12.4.0 on a 2.7GHz MacBook Pro with 16GB memory.

$u$ through $\theta$, we compute estimates $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ using the criterion functions in display (17) as this is more computationally efficient than using criterions (13) and (14). The multipliers $\lambda$ on the minimum divergence problem in (17) therefore differ from the multipliers $\lambda$ in criterions (13) and (14) by the factor $\eta$ (see the discussion in Section 2.3). As our bootstrap methods are derived based on criterions (13) and (14), when implementing the bootstrap we rescale the multiplier $\lambda$ from the minimum divergence problem in (17) by the multiplier $\eta$ on the constraint $\Delta(\theta; \gamma, \hat{P}) \leq \delta$ in the outer optimization.[28] Given the separate computations of $\eta$ (outer optimization) and $\lambda$ (inner optimization), it is computationally most convenient to implement our bootstrap CSs using $\hat{\nu} = 0$. As discussed in Section 6.2, this choice is valid but possibly conservative. Despite this potentially conservative choice, the bootstrap CSs don't seem materially wider than the bootstrap CSs under the i.i.d. Gumbel assumption.

**Computation times.** Table 6 reports computation times for the inner optimization for evaluating the criterion functions $\hat{\underline{K}}_\delta(\theta; \hat{\gamma}, \hat{P})$ and $\hat{\overline{K}}_\delta(\theta; \hat{\gamma}, \hat{P})$ at the parameter values at which $\hat{\underline{\kappa}}_\delta$ and $\hat{\overline{\kappa}}_\delta$ are attained. The computation times correspond to solving the minimum divergence problem $\Delta(\theta; \hat{\gamma}, \hat{P})$ because $k$ does not depend on $u$ (cf. display (17)). The outer optimizations were typically solved in a few minutes in a 8-core environment with 64GB memory.

**Sensitivity to $\phi$.** Implementing the procedure with $\chi^2$-divergence produced bounds that were at most 3% narrower and no wider than the bounds for hybrid divergence for all values of neighborhood size $\delta$. Repeating the analysis with $L^4$-divergence, which is stronger than $\chi^2$ and

[28] This rescaling can also be justified by applying, e.g., a suitable modification of Corollary 5 of Milgrom and Segal (2002), to derive the directional derivative of $P \mapsto \sup_{\theta \in \Theta: \delta - \Delta(\theta;P) \geq 0} k(\theta)$ analogously to Theorem 6.2. Note by similar arguments to Theorem 6.2 that the directional derivative of $P \mapsto \Delta(\theta; P)$ at $P_0$ is

$$\lim_{n\to\infty} t_n^{-1} \left(\Delta(\theta; P_0 + t_n h_n) - \Delta(\theta; P_0)\right) = \sup_{\lambda_{12} \in \underline{\Lambda}(\theta; P_0)} -\lambda_{12}' h$$

provided $\Delta(\theta; P_0) < \infty$ and Condition S' holds at $(\theta, P_0)$, where $\underline{\Lambda}(\theta; P_0)$ is constructed analogously to $\underline{\Lambda}_\delta(\theta; P)$ in Section 6.2 using the set of multipliers that solve the minimum divergence problem $\Delta(\theta; P_0)$.

hybrid divergence, produced bounds that were 15-20% narrower than the bounds for hybrid divergence for values of $\delta$ up to $\delta = 1$ and at most 5% narrower than the hybrid divergence bounds for larger values of $\delta$. As with the matching application, these results again show that the conclusions we draw from our analysis are not sensitive to the choice of $\phi$ function.

# E Proofs 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 Proofs for Section 2

**Proof of Proposition 2.1.** Immediate from Proposition G.1 in Appendix G.2. ■

Recall Condition S from Definition 2.1 and Condition $\mathrm{S}_{np}$ from Definition B.1.

**Lemma E.1** *Suppose that Assumption $\Phi$ holds and $\mu$ and $F_*$ are mutually absolutely continuous. Then Condition S holds at $(\theta, \gamma, P)$ if and only if Condition $S_{np}$ holds at $(\theta, \gamma, P)$.*

**Proof of Lemma E.1.** In view of Hölder's inequality for Orlicz classes (see (A.1)), Assumption $\Phi$ implies $\mathcal{N}_\infty = \{F : D_\phi(F \| F_*) < \infty\} \subseteq \mathcal{F}_\theta$. Therefore,

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

By Corollary 6.6.2 of Rockafellar (1970), it suffices to show $\mathrm{ri}(\mathcal{G}_\infty) = \mathrm{ri}(\mathcal{G}_\theta)$. As $\mathrm{ri}(\mathcal{G}_\infty) \subseteq \mathcal{G}_\infty \subseteq \mathcal{G}_\theta$, 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 $\mu$ are mutually absolutely continuous, $F$ has a density, say $m$, with respect to $F_*$. For each $n \geq 1$, let $m(u) \wedge n = \min\{m(u), n\}$ and define

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

Each $F_n$ defined by $\mathrm{d}F_n = m_n \mathrm{d}F_*$ belongs to $\mathcal{N}_\infty$. It follows that $\mathbb{E}^{F_n}[g(U, \theta, \gamma)] \in \mathcal{G}_\infty$. By monotone convergence, we have $\mathbb{E}^{F_n}[g(U, \theta, \gamma)] \to x$. 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. 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 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 $(\theta, \gamma_0, P_0)$ 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 $(\theta, \gamma_0, P_0)$ either because $\mathcal{N}_\infty \subseteq \mathcal{F}_\theta$ for all $\theta$ under Assumption $\Phi$. Therefore, $\underline{K}_\infty(\theta; \gamma_0, P_0) = +\infty$ in that case too. We therefore have

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

In view of Lemma B.1, it suffices to show $\inf \mathcal{K} = \inf \mathcal{K}_\infty$. Note that $\inf \mathcal{K} \leq \inf \mathcal{K}_\infty$ holds by virtue of the inclusion $\mathcal{N}_\infty \subseteq \mathcal{F}_\theta$ for all $\theta$. For the reverse inequality, choose any $\epsilon > 0$. By $S$-regularity of $\Theta_I$, there exists $\underline{\theta} \in \Theta_I$ for which Condition S holds at $(\underline{\theta}, \gamma_0, P_0)$ and for which $\underline{K}_{np}(\underline{\theta}; \gamma_0, P_0) \leq \inf \mathcal{K} + \epsilon$. As Condition S holds at $(\underline{\theta}, \gamma_0, P_0)$ and $\mu \ll F_* \ll \mu$, Lemma E.1 implies that Condition $\text{S}_{np}$ must also hold at $(\underline{\theta}, \gamma_0, P_0)$. Moreover, the $\mu$-essential infimum and $F_*$-essential infimum of any function are equal because $\mu \ll F_* \ll \mu$. Therefore by Lemmas B.2 and B.3, we have $\underline{K}_\infty(\underline{\theta}; \gamma_0, P_0) = \underline{K}_{np}(\underline{\theta}; \gamma_0, P_0)$. It follows by definition of $\inf \mathcal{K}_\infty$ that $\inf \mathcal{K}_\infty \leq \underline{K}_\infty(\underline{\theta}; \gamma_0, P_0) = \underline{K}_{np}(\underline{\theta}; \gamma_0, P_0) \leq \inf \mathcal{K} + \epsilon$. Therefore, $\inf \mathcal{K}_\infty \leq \inf \mathcal{K}$. ■

## E.2 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

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

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

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

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

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

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

Note $v^B \leq v^A$ 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

$$\inf_{F \in \mathcal{N}_\delta} 0 \quad \text{subject to (19) and } \mathbb{E}^F[k(U, \theta, \gamma)] = \kappa \text{ holding at } (\theta, \gamma, P). \tag{49}$$

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

$$\mathbb{E}^{F_\varepsilon^A}[k(U, \theta_\varepsilon^A, \gamma)] \leq v^A + \varepsilon\,.$$

Then $\underline{F}_\varepsilon^A$ solves the feasibility program (49) with $\theta = \theta_\varepsilon^A$ and $\kappa = \kappa_\varepsilon^A := \mathbb{E}^{F_\varepsilon^A}[k(U, \theta_\varepsilon^A, \gamma)]$. Note that $\mathbb{E}^{F_\varepsilon^A}[g_{4e}(U, \theta_\varepsilon^A, \gamma)] = 0$ also holds by construction. Therefore, $(\kappa_\varepsilon^A, \theta_\varepsilon^A)$ are feasible for the augmented form of Program B. It follows that $v^B \leq \kappa_\varepsilon^A \leq v^A + \varepsilon$ holds for each $\varepsilon > 0$. As $\varepsilon > 0$ is arbitrary, we have $v^B \leq v^A$ whenever $v^A > -\infty$.

A similar argument applies if $v^A = -\infty$: for any $n \in \mathbb{N}$, we may choose $\theta_n^A \in \Theta$ and $F_n^A \in \mathcal{N}_\delta$ such that the constraints in Program A are satisfied and $\mathbb{E}^{F_n^A}[k(U, \theta_n^A, \gamma)] \leq -n$. It follows that $v^B \leq -n$. Hence, $v^B = v^A = -\infty$. ■

**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 $k$ and $g$ on $(\theta, \gamma)$ to simplify notation in what follows.

First, suppose $k$ depends on $u$. The dual formulation is justified by Proposition 2.1. A dual solution $(\underline{\eta}, \underline{\zeta}, \underline{\lambda})$ exists by Proposition G.1(iii).

Suppose $\underline{\eta} > 0$. We wish to show that the change of measure $\underline{m}_{\delta,\theta}(u) = \dot{\phi}^\star(-\underline{\eta}^{-1}(k(u) + \underline{\zeta} + \underline{\lambda}' g_s(u)))$ induces a distribution that solves the primal problem (20) at $\theta$. Differentiability of the objective function in $(\eta, \zeta, \lambda)$ is guaranteed by Assumption $\Phi$. Also note that Assumption $\Phi$(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 $(\theta,\gamma,P)$ 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 $\Phi$(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 suppose $\underline{\eta} = 0$. Here we wish to show that $\underline{m}_{\delta,\theta}(u) = 1\!\!1\{u \in \underline{A}_{\delta,\theta}\}/F_*(\underline{A}_{\delta,\theta})$ induces a distribution that solves the primal problem (20) at $\theta$. As the neighborhood constraint $F \in \mathcal{N}_\delta$ is not binding, the value of the objective must be the same as the optimal value when $\delta = \infty$. In view of Lemma B.2, the value is $F_*\text{-ess}\inf(k(\cdot) + \underline{\lambda}'g_s(\cdot)) - \underline{\lambda}_{12}'P$. We can write problem (22) as a nested optimization:

$$\sup_{\lambda\in\Lambda_s}\left(\sup_{\eta>0,\zeta\in\mathbb{R}} -\eta\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U)+\zeta+\lambda'g_s(U)}{-\eta}\right)\right] - \eta\delta - \zeta - \lambda_{12}'P\right).$$

At $\lambda = \underline{\lambda}$, the inner problem is the dual of $\inf_{F\in\mathcal{N}_\delta}\mathbb{E}^F[k(U) + \underline{\lambda}'g_s(U) - \underline{\lambda}_{12}'P]$. As $\underline{\eta} = 0$, the constraint $F \in \mathcal{N}_\delta$ is not binding and so the minimizing distribution must be supported on $\underline{A}_{\delta,\theta}$. Finally, by convexity of $\phi$, the distribution induced by $\underline{m}_{\delta,\theta}$ minimizes $D_\phi(\,\cdot\,\|F_*)$ among all distributions with support $\underline{A}_{\delta,\theta}$.

Now suppose $k$ does not depend on $u$. By Proposition G.2, the primal and dual values of (15) are equal and a dual solution exists. By similar arguments to above, $\mathbb{E}^{F_*}[\underline{m}_{\delta,\theta}(U)] = 1$, and (1a)–(1c) and $\mathbb{E}^F[g_{4s}(U,\theta,\gamma)] = 0$ hold at $(\theta,\gamma,P)$ under $\underline{F}_{\delta,\theta}$. Finally, as there exists $F \in \mathcal{N}_\delta$ under which the moment conditions (1a)–(1c) and $\mathbb{E}^F[g_{4s}(U,\theta,\gamma)] = 0$ hold at $(\theta,\gamma,P)$, we must have $D(\underline{F}_{\delta,\theta}\|F_*) \le D(F\|F_*) \le \delta$, as required. ■

## E.3 Proofs for Section 4

**Proof of Proposition 4.1.** As $\phi_1(x) \le \bar{a}\phi_2(x)$ for all $x > 0$, we have $D_{\phi_1}(F\|F_*) \le \bar{a}D_{\phi_2}(F\|F_*)$. Hence, $\mathcal{N}_{\delta,2} \subseteq \mathcal{N}_{\bar{a}\delta,1}$ for each $\delta > 0$. The result follows from this inclusion, noting that $\underline{\kappa}_{\bar{a}\delta,1}$ and $\underline{\kappa}_{\bar{a}\delta,1}$ are both finite because Assumption $\Phi$ holds for $\phi_1$. ■

## E.4 Proofs for Section 6

We first present some preliminary lemmas.

**Lemma E.2** *Suppose that Assumptions $\Phi$ and M(i),(v) hold. Let $\{(F_n,\theta_n,\gamma_n,P_n)\} \subseteq \mathcal{N}_\delta \times \Theta \times \Gamma \times \mathcal{P}$ with $(\gamma_n,P_n) \to (\tilde{\gamma},\tilde{P}) \in \Gamma \times \mathcal{P}$ and with (1) holding under $F_n$ at $(\theta_n,\gamma_n,P_n)$. 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 $\lim_{l\to\infty}\mathbb{E}^{F_{n_l}}[k(U,\theta_{n_l},\gamma_{n_l})] = \mathbb{E}^{\tilde{F}}[k(U,\tilde{\theta},\tilde{\gamma})]$ and similarly for each entry of $g_1,\dots,g_4$, and (1) holds under $\tilde{F}$ at $(\tilde{\theta},\tilde{\gamma},\tilde{P})$.*

**Proof of Lemma E.2.** Let $m_n = \frac{\mathrm{d}F_n}{\mathrm{d}F_*}$. By Assumption M(v), $\{\theta_n\}$ has a convergent subsequence $\{\theta_{n_l}\}$. As $\{m_{n_l}\}$ is $\|\cdot\|_\phi$-norm bounded (Lemma F.1(ii)), taking a further subsequence if necessary we may assume $\{m_{n_l}\}$ is $\mathcal{E}$-weakly convergent to $\tilde{m} \in \mathcal{L}$ (see Appendix F). By the triangle inequality, the Hölder inequality (A.1), $\mathcal{E}$-weak convergence, and Assumption M(i), we have

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

It follows by similar arguments that

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

Finally, By Lemma F.1(i), we have $\delta \ge \liminf_{l\to\infty}\mathbb{E}^{F_*}[\phi(m_{n_l}(U))] \ge \mathbb{E}^{F_*}[\phi(\tilde{m}(U))]$. ■

**Lemma E.3** *Suppose that Assumptions* $\Phi$ *and* M*(i),(iii)–(v) hold. Then* $\underline{\kappa}_\delta$ *and* $\overline{\kappa}_\delta$ *are finite, and*

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

**Proof of Lemma E.3.** We prove the result only for $\underline{\kappa}_\delta$; the result for $\overline{\kappa}_\delta$ follows similarly.

Finiteness of $\underline{\kappa}_\delta$ follows by Assumptions $\Phi$ and M(i)(v) and the Hölder inequality (A.1). To simplify notation, we suppress dependence of $\Theta_\delta(\gamma_0,P_0)$ on $(\gamma_0,P_0)$ 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 (1) at $(\underline{\theta},\gamma_0,P_0)$. As $\Delta(\underline{\theta};\gamma_0,P_0) = \delta$, it follows by convexity of $\phi$ that $F_{\underline{\theta}}$ must be unique. Therefore

$$\mathbb{E}^{F_{\underline{\theta}}}[k(U,\underline{\theta},\gamma_0)] = \underline{K}_\delta(\underline{\theta};\gamma_0,P_0) < \inf_{\theta\in\Theta_\delta} \underline{K}_\delta(\theta;\gamma_0,P_0) \le \inf_{\theta\in\Theta_\delta} \mathbb{E}^{F_\theta}[k(U,\theta_0,\gamma_0)]\,, \tag{50}$$

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

Choose $\{\theta_n\} \subset \Theta_\delta$ with $\theta_n \to \underline{\theta}$ (we may choose such a sequence by Assumption M(iv)). By Lemma E.2, there is a subsequence $\{(\theta_{n_l}, F_{\theta_{n_l}}, \gamma_0, P_0)\}$ 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 $F_{\underline{\theta}}$ that $\underline{F} = F_{\underline{\theta}}$. By Lemma E.2, 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 (50). ■

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.4** *Suppose that Assumptions* $\Phi$ *and* M*(i)–(v) hold. Then* $\underline{b}_\delta(\gamma,P)$ *and* $\overline{b}_\delta(\gamma,P)$ *are continuous at* $(\gamma_0,P_0)$*.*

**Proof of Lemma E.4.** We prove the result only for $\underline{b}_\delta$; the result for $\overline{b}_\delta$ follows similarly.

Fix $\varepsilon > 0$. By Lemma E.3, 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 G.8 and Assumption M(ii) we have $\Delta(\theta_\varepsilon;\gamma,P) < \delta$ on a neighborhood $N$ of $(\gamma_0,P_0)$. Moreover, by Lemma G.9(i) and Assumption M(i)–(iii) we have

$$\underline{K}_\delta(\theta_\varepsilon;\gamma,P) < \underline{K}_\delta(\theta_\varepsilon;\gamma_0,P_0) + \varepsilon$$

on a neighborhood $N'$ of $(\gamma_0, P_0)$. On $N \cap N'$ we therefore have

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

establishing u.s.c. of $\underline{b}_\delta(\gamma, P)$ at $(\gamma_0, P_0)$.

To establish l.s.c., suppose there is $\varepsilon > 0$ and $(\gamma_n, P_n) \to (\gamma_0, P_0)$ along which

$$\underline{b}_\delta(\gamma_n, P_n) \le \underline{b}_\delta(\gamma_0, P_0) - 2\varepsilon \,. \tag{51}$$

Note $\Theta_\delta(\gamma_n, P_n)$ is nonempty for $n$ sufficiently large by Lemma G.8 and Assumption 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

$$\mathbb{E}^{F_n}[k(U, \theta_n, \gamma_n)] < \underline{b}_\delta(\gamma_n, P_n) + \varepsilon \,. \tag{52}$$

By Lemma E.2 there is a subsequence $(F_{n_l}, \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \to (\underline{F}, \underline{\theta}, \gamma_0, P_0)$ 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

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

In view of (51) and (52) and Lemma E.3, this implies $\underline{K}_\delta(\underline{\theta}; \gamma_0, P_0) \le \underline{b}_\delta(\gamma_0, P_0) - \varepsilon = \underline{\kappa}_\delta - \varepsilon$, contradicting the definition of $\underline{\kappa}_\delta$. ∎

**Proof of Theorem 6.1.** Note that $\underline{\kappa}_\delta = \underline{b}_\delta(\gamma_0, P_0)$ and $\overline{\kappa}_\delta = \overline{b}_\delta(\gamma_0, P_0)$ by Lemma E.3 and $\hat{\underline{\kappa}}_\delta = \underline{b}_\delta(\hat{\gamma}, \hat{P})$ and $\hat{\overline{\kappa}}_\delta = \overline{b}_\delta(\hat{\gamma}, \hat{P})$ by definition. The result now follows by Lemma E.4 and Slutsky's theorem. ∎

**Lemma E.5** *Suppose that Assumptions $\Phi$ and M(i),(ii) hold, Condition S' holds at $(\theta, \gamma, P)$, and $\Delta(\theta; \gamma, P) < \delta$. Then there is a neighborhood $N$ of $(\theta, \gamma, P)$ such that Condition S' holds at $(\tilde{\theta}, \tilde{\gamma}, \tilde{P})$ and $\Delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P}) < \delta$ for all $(\tilde{\theta}, \tilde{\gamma}, \tilde{P}) \in N$.*

**Proof of Lemma E.5.** By Lemma G.7, Condition S' holds at all $(\tilde{\theta}, \tilde{\gamma}, \tilde{P})$ in a neighborhood $N'$ of $(\theta, \gamma, P)$. Moreover, $\Delta(\tilde{\theta}; \tilde{\gamma}, \tilde{P}) < \delta$ holds at all $(\tilde{\theta}, \tilde{\gamma}, \tilde{P})$ in a neighborhood $N''$ of $(\theta, \gamma, P)$ by Lemma G.8. Set $N = N' \cap N''$. ∎

In the remainder of this subsection we drop dependence of all quantities on $\gamma$.

**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 $\{\theta_n\}$ such that $\underline{K}_\delta(\theta_n; P_0) \downarrow \underline{\kappa}_\delta$. Let $F_n$ solve the primal problem for $\theta_n$. By Lemma E.2, there is a subsequence $(F_{n_l}, \theta_{n_l}) \to (\underline{F}, \underline{\theta})$ with $\underline{F} \in \mathcal{N}_\delta$ and $\underline{\theta} \in \Theta$ 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. We may deduce by similar arguments that $\underline{\Theta}_\delta(P_0)$ is closed. Compactness now follows by Assumption M(v).

Step 2: We now prove directional differentiability. Let $P_n = P_0 + t_n h_n$ with $t_n \downarrow 0$ and $h_n \to h$. Choose $\underline{\theta} \in \underline{\Theta}_\delta(P_0)$. By Lemma E.5 and Assumption M(iii)(vi), Condition S' holds at $(\underline{\theta}, P_n)$ and $\Delta(\underline{\theta}; P_n) < \delta$ for $n$ sufficiently large, so by Proposition G.1(iv) the set $\underline{\Lambda}_\delta(\underline{\theta}; P_n)$ is nonempty and compact for $n$ sufficiently large. It now follows by definition of the objective (13) that

$$\underline{b}_\delta(P_n) - \underline{b}_\delta(P_0) \le \underline{K}_\delta(\underline{\theta}; P_n) - \underline{K}_\delta(\underline{\theta}; P_0) \le t_n \times -\underline{\lambda}_{12}' h_n\,,$$

for all $\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}; P_n)$. Finally, by Lemma G.9(ii) we obtain

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

Taking the infimum of both sides over $\underline{\theta} \in \underline{\Theta}_\delta$ yields

$$\limsup_{n\to\infty} \frac{\underline{b}_\delta(P_n) - \underline{b}_\delta(P_0)}{t_n} \le \inf_{\theta\in\underline{\Theta}_\delta} \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\theta;P_0)} -\underline{\lambda}_{12}' h\,. \tag{53}$$

For the lower bound, choose $\theta_n \in \Theta_\delta(P_n)$ with $\underline{K}_\delta(\theta_n; P_n) \le \underline{b}_\delta(P_n) + t_n^2$ for all $n$ sufficiently large. Take a subsequence $\{\theta_{n_l}\}$. By Assumption M(v) (taking a further subsequence if necessary), we have $\theta_{n_l} \to \underline{\theta} \in \Theta$. By similar arguments to step 1, we may in fact deduce that $\underline{\theta} \in \underline{\Theta}_\delta$. Reasoning as above, for $l$ sufficiently large we have

$$\underline{b}_\delta(P_{n_l}) - \underline{b}_\delta(P_0) \ge \underline{K}_\delta(\theta_{n_l}; P_{n_l}) - \underline{K}_\delta(\theta_{n_l}; P_0) - t_{n_l}^2 \ge t_{n_l} \times -\underline{\lambda}_{12}' h_{n_l} - t_{n_l}^2\,,$$

where the final inequality holds for any $\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta_{n_l}; P_0)$. By Assumption M(vii), we may choose $\underline{\lambda}_{12,n_l} \in \underline{\Lambda}_\delta(\theta_{n_l}; P_0)$ for which $-\underline{\lambda}_{12,n_l}' h \to \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\underline{\theta};P_0)} -\underline{\lambda}_{12}' h$ as $l \to \infty$. Therefore,

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

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

$$\liminf_{n\to\infty} \frac{\underline{b}_\delta(P_n) - \underline{b}_\delta(P_0)}{t_n} \ge \inf_{\theta\in\underline{\Theta}_\delta} \max_{\underline{\lambda}_{12}\in\underline{\Lambda}_\delta(\theta;P_0)} -\underline{\lambda}_{12}' h\,, \tag{54}$$

proving directional differentiability. Finally, Assumption M(vii) and Lemma G.9(ii) imply $\theta \mapsto \underline{\Lambda}_\delta(\theta; P_0)$ is continuous at each $\theta \in \underline{\Theta}_\delta$. The set $\underline{\Lambda}_\delta(\theta; P_0)$ is also compact for each $\theta \in \underline{\Theta}_\delta$ by Proposition G.1(iv). It follows by the maximum theorem that the infima in (53) and (54) can be replaced by minima.

Step 3: In view of step 2, the asymptotic distribution follows by Theorem 2.1 of Shapiro (1991) and the fact that $\sqrt{n}(\hat{P} - P) \to_d N(0, \Sigma)$. ■

**Lemma E.6** *Let $\underline{\Theta}_{\delta,a} = \{\theta \in \Theta_\delta(P_0) : \underline{K}_\delta(\theta; P_0) \leq \underline{b}_\delta(P_0) + a\}$. Suppose that Assumptions $\Phi$ and M(i)-(iii),(v),(vi) hold. Then there is $a > 0$ and a neighborhood $N$ of $P_0$ such that Condition S' holds at $(\theta, P)$ and $\Delta(\theta; P) < \delta$ for all $(\theta, P) \in \underline{\Theta}_{\delta,a} \times N$.*

**Proof of Lemma E.6.** Suppose the assertion is false. Then there is $a_n \downarrow 0$, $\theta_n \in \underline{\Theta}_{\delta,a_n}$, and $P_n \to P$ for which Condition S' does not hold at $(\theta_n; P_n)$ and/or $\Delta(\theta_n; P_n) \geq \delta$. By Assumption M(v), we can extract a convergent subsequence $\theta_{n_l} \to \underline{\theta}$. By similar arguments to Step 1 of the proof of Theorem 6.2, we may deduce that $\underline{\theta} \in \underline{\Theta}_\delta$. Then by Lemma E.5 and Assumption M(iii)(vi) we must have that Condition S' holds at $(\theta_{n_l}, P_{b_l})$ and $\Delta(\theta_{n_l}; P_{n_l}) < \delta$ for all $l$ sufficiently large, a contradiction. ■

**Lemma E.7** *Suppose that the conditions of Theorem 6.3 hold. Then:*

*(i) $|\widehat{d\underline{b}}_{\delta,P_0}[h_1] - \widehat{d\underline{b}}_{\delta,P_0}[h_2]| \leq \underline{C}_n \|h_1 - h_2\|$ and $|\widehat{d\overline{b}}_{\delta,P_0}[h_1] - \widehat{d\overline{b}}_{\delta,P_0}[h_2]| \leq \overline{C}_n \|h_1 - h_2\|$ 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.7.** We prove the results for the lower values; the results for the upper values follow similarly.

Step 1: Recall $\underline{\Theta}_{\delta,a}$ from Lemma E.6. We show there exists a sequence $a_n \downarrow 0$ such that $\underline{\Theta}_\delta \subseteq \hat{\underline{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a_n}$ holds with probability approaching one (wpa1).

Fix any sufficiently small $a > 0$ and let $N$ be the neighborhood from Lemma E.6. As $\underline{\Theta}_\delta \subset \underline{\Theta}_{\delta,a}$, for all $(\theta, P) \in \underline{\Theta}_\delta \times N$, we have that $\Delta(\theta; P) < \delta$ and Condition S' holds at $(\theta, P)$. Therefore, $\underline{\Theta}_\delta \subseteq \Theta_\delta(\hat{P})$ wpa1 by consistency of $\hat{P}$. Moreover, $\underline{\Lambda}_\delta(\theta; P)$ is nonempty and compact for all $(\theta, P) \in \underline{\Theta}_\delta \times N$ (cf. Proposition G.1(iv)). By similar arguments to the proof of Theorem 6.2, we have

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

for any $(\theta, P) \in \underline{\Theta}_\delta \times N$, and so

$$\sup_{\theta \in \underline{\Theta}_\delta} \underline{K}_\delta(\theta; P) - \underline{b}_\delta(P) \leq \sup_{\theta \in \underline{\Theta}_\delta} \max_{\underline{\lambda}_{12} \in \Lambda_\delta(\theta; P)} -\underline{\lambda}_{12}'(P - P_0) + \underline{b}_\delta(P_0) - \underline{b}_\delta(P). \quad (55)$$

As $\underline{\Theta}_\delta$ is compact (see Step 1 of the proof of Theorem 6.2) and $(\theta, P) \mapsto \max_{\underline{\lambda}_{12} \in \Lambda_\delta(\theta;P)} \|\underline{\lambda}_{12}\|$ is u.s.c. (by Lemma G.9(ii)) on $\underline{\Theta}_\delta \times N$, we may choose a neighborhood $N'$ of $P_0$ upon which $\sup_{\theta \in \underline{\Theta}_\delta} \max_{\underline{\lambda}_{12} \in \Lambda_\delta(\theta;P)} \|\underline{\lambda}_{12}\| < \infty$. Setting $P = \hat{P}$ in (55) and noting $\underline{b}_\delta(P_0) - \underline{b}_\delta(\hat{P}) = O_p(n^{-1/2})$ by Theorem 6.2, we may deduce that $\sup_{\theta \in \underline{\Theta}_\delta} \underline{K}_\delta(\theta; \hat{P}) - \underline{b}_\delta(\hat{P}) \le O_p(n^{-1/2})$ and therefore that $\underline{\Theta}_\delta \subseteq \hat{\underline{\Theta}}_{\delta,n}$ wpa1.

By the almost sure representation theorem (see, e.g., Shapiro (1991, Theorem A1)), there exists a sequence of random vectors $\{(Z_n, \hat{\nu}_n)\}$ and a random vector $Z$ defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ 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 $(Z_n(\omega), \hat{\nu}_n(\omega)) \to (Z(\omega), \nu(\omega))$. Let $\hat{\underline{\Theta}}_{\delta,n}(\omega) = \{\theta \in \Theta_\delta(P_n(\omega)) : \underline{K}_\delta(\theta; P_n(\omega)) \le \underline{b}_\delta(P_n(\omega)) + \nu(\omega)\sqrt{\log n / n}\}$. The set $\Theta_\delta(P_n(\omega))$, and therefore $\hat{\underline{\Theta}}_{\delta,n}(\omega)$, is nonempty for $n$ sufficiently large.

Suppose there is $\{\theta_{n_l}(\omega)\}$ with $\theta_{n_l}(\omega) \in \hat{\underline{\Theta}}_{\delta,n}(\omega) \setminus \underline{\Theta}_{\delta,a}$ for all $l$. By Assumption M(v) (taking a further subsequence if necessary) we have $\theta_{n_l}(\omega) \to \underline{\theta}(\omega) \in \Theta$. We may deduce by similar arguments to Step 1 in the proof of Theorem 6.2 that $\underline{\theta}(\omega) \in \underline{\Theta}_\delta$. Then by Lemma G.9(i) and Assumption M(iii)(vi), we have $\lim_{l \to \infty} \underline{K}_\delta(\theta_{n_l}(\omega); P_0) = \underline{K}_\delta(\underline{\theta}(\omega); P_0) = \underline{b}_\delta(P_0)$. Moreover, $\lim_{l \to \infty} \Delta(\theta_{n_l}(\omega); P_0) = \Delta(\underline{\theta}(\omega); P_0) < \delta$ by Lemma G.8 and Assumption M(iii)(vi). Therefore, $\theta_{n_l}(\omega) \in \underline{\Theta}_{\delta,a}$ for $l$ sufficiently large, a contradiction. Therefore, $\hat{\underline{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a}$ wpa1 for each $a > 0$.

In view of the above, we may choose $a_n \downarrow 0$ so that $\hat{\underline{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a_n}$ wpa1.

Step 2: We prove part (i). Let $a > 0$ and the neighborhood $N$ be as in Lemma E.6. It follows by Proposition G.1(iv) that $\underline{\Lambda}_\delta(\theta; P)$ is compact and nonempty for all $(\theta, P) \in \underline{\Theta}_{\delta,a} \times N$. We may also deduce by similar arguments to Step 1 in the proof of Theorem 6.2 that $\underline{\Theta}_{\delta,a}$ is compact. Finally, as $(\theta, P) \mapsto \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta;P)} \|\underline{\lambda}_{12}\|$ is u.s.c. (by Lemma G.9(ii)) on $\underline{\Theta}_{\delta,a} \times N$, we have $\sup_{\theta \in \underline{\Theta}_{\delta,a}} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta;P)} \|\underline{\lambda}_{12}\| \le C$ on a neighborhood $N'$ of $P_0$ for some $C < \infty$. Now, as $\hat{P} \in N \cap N'$ and $\hat{\underline{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a_n}$ both hold wpa1, it follows from the fact that the $\max$ and $\min$ operations are Lipschitz and the Cauchy-Schwarz inequality that

$$\left| \widehat{d\underline{b}}_{\delta,P_0}[h_1] - \widehat{d\underline{b}}_{\delta,P_0}[h_2] \right| \le C\|h_1 - h_2\| \quad \text{for all } h_1, h_2 \in \mathbb{R}^{d_1+d_2}$$

wpa1, proving part (i).

Step 3: We prove part (ii). As $\underline{\Theta}_\delta \subseteq \hat{\underline{\Theta}}_{\delta,n} \subseteq \underline{\Theta}_{\delta,a_n}$ wpa1 by step 1, it suffices to show

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

for all $h \in \mathbb{R}^{d_1+d_2}$. Using the almost sure representation in 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 G.9(ii), we have

$$\begin{aligned}
&\limsup_{n\to\infty} \inf_{\theta \in \underline{\Theta}_\delta} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta, P_n(\omega))} -\underline{\lambda}_{12}' h \\
&\le \limsup_{n\to\infty} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}_h, P_n(\omega))} -\underline{\lambda}_{12}' h \le \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}_h, P_0)} -\underline{\lambda}_{12}' h = d\underline{b}_{\delta, P_0}[h] \,. \qquad (56)
\end{aligned}$$

Now choose $\underline{\theta}_n(\omega) \in \underline{\Theta}_{\delta, a_n}$ for which

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

Take any subsequence $\{\theta_{n_l}(\omega), P_{n_l}(\omega)\}$. By Assumption M(v) (taking a further subsequence if necessary) we have $\theta_{n_l}(\omega) \to \underline{\theta}(\omega) \in \Theta$. By similar arguments to Step 1 of the proof of Theorem 6.2, we may deduce $\underline{\theta}(\omega) \in \underline{\Theta}_\delta$. Lemma G.9(ii) and Assumption M(vii') together imply the correspondence $(\theta, P) \mapsto \underline{\Lambda}_\delta(\theta, P)$ is continuous at $(\underline{\theta}, P_0)$. Therefore,

$$\begin{aligned}
&\liminf_{l\to\infty} \inf_{\theta \in \underline{\Theta}_{\delta, a_{n_l}}} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta, P_{n_l}(\omega))} -\underline{\lambda}_{12}' h \\
&\ge \liminf_{l\to\infty} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta_{n_l}(\omega), P_{n_l}(\omega))} -\underline{\lambda}_{12}' h \ge \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\underline{\theta}(\omega), P_0)} -\underline{\lambda}_{12}' h \ge d\underline{b}_{\delta, P_0}[h] \,.
\end{aligned}$$

As the lower bound on the right-hand side does not depend on the subsequence, we therefore have $\liminf_{n\to\infty} \inf_{\theta \in \underline{\Theta}_{\delta, a_n}} \max_{\underline{\lambda}_{12} \in \underline{\Lambda}_\delta(\theta, P_n(\omega))} -\underline{\lambda}_{12}' h \ge d\underline{b}_{\delta, P_0}[h]$. This, in conjunction with the upper bound (56), 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 hold by Theorem 6.2 and because $\sqrt{n}(\hat{P} - P_0) \to_d N(0, \Sigma)$ with $\Sigma$ finite, respectively. Their Assumption 3 is assumed directly. Finally, Lemma E.7 shows that $\widehat{d\underline{b}}_{\delta, P_0}$ and $\widehat{d\overline{b}}_{\delta, P_0}$ satisfy the sufficient conditions for their Assumption 4 presented in their Remark 3.4. This proves consistency. Coverage of $CS_{\delta,L}^{1-\alpha}$ and $CS_{\delta,U}^{1-\alpha}$ follows by continuity of the distribution functions. Coverage of $CS_\delta^{1-\alpha}$ follows by the Bonferroni inequality. ■

**Proof of Theorem 6.4.** We prove the result only for $CS_\delta^{1-\alpha}$; the result for the other CSs follow similarly. Say that $P_0 \in CS_{P_0}^{1-\alpha}$ if $P_{10} \le \hat{P}_{1,U}^{1-\alpha}$ and $P_{20} \in [\hat{P}_{2,L}^{1-\alpha}, \hat{P}_{2,U}^{1-\alpha}]$ both hold. By Lemma E.3, for each $\varepsilon > 0$ we may choose $\underline{\theta}_\varepsilon, \overline{\theta}_\varepsilon \in \Theta_\delta(P_0)$ such that $\underline{K}_\delta(\underline{\theta}_\varepsilon; P_0) < \underline{\kappa}_\delta + \varepsilon$ and $\overline{K}_\delta(\overline{\theta}_\varepsilon; P_0) > \overline{\kappa}_\delta - \varepsilon$. Let $\underline{F}_{\underline{\theta}_\varepsilon}$ and $\overline{F}_{\overline{\theta}_\varepsilon}$ solve problem (15) at $(\underline{\theta}_\varepsilon; P_0)$ and $(\overline{\theta}_\varepsilon; P_0)$, respectively. Whenever $P_0 \in CS_{P_0}^{1-\alpha}$ holds, $\underline{F}_{\underline{\theta}_\varepsilon}$ and $\overline{F}_{\overline{\theta}_\varepsilon}$ must also satisfy the "relaxed" moment conditions used for computing $\hat{\underline{\kappa}}_{\delta,1-\alpha}$ and $\hat{\overline{\kappa}}_{\delta,1-\alpha}$, so it follows that $\Delta_{cs}(\underline{\theta}_\varepsilon; \hat{P}_{1-\alpha}) < \delta$ and $\Delta_{cs}(\overline{\theta}_\varepsilon; \hat{P}_{1-\alpha}) < \delta$. Moreover, as the primal solutions for $\underline{K}_\delta(\underline{\theta}_\varepsilon; P_0)$ and $\overline{K}_\delta(\overline{\theta}_\varepsilon; P_0)$ are feasible for the relaxed problem whenever $P_0 \in CS_{P_0}^{1-\alpha}$, we have

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

and similarly $\hat{\overline{\kappa}}_{\delta,1-\alpha} > \overline{\kappa}_\delta - \varepsilon$. As $\varepsilon$ is arbitrary, we have that $\underline{\kappa}_\delta \geq \underline{\hat{\kappa}}_{\delta,1-\alpha}$ and $\overline{\kappa}_\delta \leq \hat{\overline{\kappa}}_{\delta,1-\alpha}$ holds whenever $P_0 \in CS_{P_0}^{1-\alpha}$. The desired coverage now follows by (30). ■

# References


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.

Bonhomme, S. and M. Weidner (2021). Minimizing sensitivity to model misspecification. *Quantitative Economics*, forthcoming.

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

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.

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2 ed.). Springer-Verlag New York, Inc.

Milgrom, P. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *Econometrica 70*(2), 583–601.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research 30*(1), 169–186.

# Secondary Online Appendix to "Counterfactual Sensitivity and Robustness"

Timothy Christensen    Benjamin Connault

(This secondary online appendix is not intended for publication.)

# F Background Material on Orlicz Spaces

In this Appendix we briefly review some relevant aspects of the theory of paired Orlicz spaces. We refer the reader to Krasnosel'skii and Rutickii (1961; KR hereafter) for a textbook treatment.

Let $L^1(F_*)$ denote the space of (equivalence classes of) all measurable $f : \mathcal{U} \to \mathbb{R}$ with finite first moment under $F_*$. Define

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

where $\psi(x) = \phi^\star(x) - x$ with $\phi^\star$ denoting the convex conjugate of $\phi$. The space $\mathcal{L}$ corresponds to the space $L^\star_M$ in KR's notation with $M(x) = \phi(1 + x)$ and $\mathcal{E}$ corresponds to $E_N$ in KR's notation with $N(x) = \psi(x)$. The condition $\lim_{x\to\infty} x\phi'(x)/\phi(x) < \infty$ in Assumption $\Phi$(i) implies $\phi(1 + x)$ satisfies KR's $\Delta_2$-condition (KR, Theorem 4.1). As such, $\mathcal{L}$ and $\mathcal{E}$ are separable Banach spaces when equipped with the Orlicz norms

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

respectively (see KR, Section 10 for a discussion of separability and KR, Theorem 10.5 for the norm). Given $\phi_1, \phi_2$ satisfying Assumption $\Phi$(i), write $\phi_1 \prec \phi_2$ if there exist positive constants $c$ and $x_0$ such that $\phi_1(x) \le \phi_2(cx)$ for all $x \ge x_0$. If $\phi_1 \prec \phi_2$ and $\phi_2 \prec \phi_1$ then $\phi_1$ and $\phi_2$ are said to be *equivalent*. Equivalent $\phi$ functions induce equivalent spaces $\mathcal{L}$ and $\mathcal{E}$ and equivalent Orlicz norms on these spaces (KR, Section 13). For example, the functions inducing hybrid and $\chi^2$ divergence are equivalent, and their spaces $\mathcal{L}$ and $\mathcal{E}$ are equivalent to $L^2(F_*)$, and the Orlicz norms $\|\cdot\|_\phi$ and $\|\cdot\|_\psi$ are equivalent to the $L^2(F_*)$ norm. Similarly, any $\phi$ that is equivalent to $x^p$ ($p > 1$) induces a space $\mathcal{L}$ equivalent to $L^p(F_*)$, an Orlicz norm $\|\cdot\|_\phi$ equivalent to the $L^p(F_*)$ norm, a space $\mathcal{E}$ equivalent to $L^q(F_*)$ with $\frac{1}{p} + \frac{1}{q} = 1$, and a norm $\|\cdot\|_\psi$ equivalent to the $L^q(F_*)$ norm. As with $L^p$ spaces, there is a version of Hölder's inequality, namely

$$|\mathbb{E}^{F_*}[f(U)g(U)]| \le \|f\|_\phi \|g\|_\psi, \tag{A.1}$$

which holds for each $f \in \mathcal{L}$ and $g \in \mathcal{E}$ (KR, Theorem 9.3).

The spaces $\mathcal{L}$ and $\mathcal{E}$ are paired spaces under the map $\langle \cdot, \cdot \rangle : \mathcal{L} \times \mathcal{E} \to \mathbb{R}$ given by

$$\langle f, g \rangle = \mathbb{E}^{F_*}[f(U)g(U)] \,.$$

A sequence $\{f_n\} \subset \mathcal{L}$ is said to be *$\mathcal{E}$-weakly convergent* if $\{\langle f_n, g \rangle\}$ converges for each $g \in \mathcal{E}$. The space $\mathcal{L}$ is $\mathcal{E}$-weakly complete: any $\mathcal{E}$-weakly convergent sequence $\{f_n\} \subset \mathcal{L}$ has a unique limit $f_0 \in \mathcal{L}$ for which

$$\lim_{n\to\infty} \langle f_n, g \rangle = \langle f_0, g \rangle$$

for each $g \in \mathcal{E}$ (KR, Theorem 14.4). The space $\mathcal{L}$ is also *$\mathcal{E}$-weakly compact*: every $\|\cdot\|_\phi$-norm bounded sequence in $\mathcal{L}$ has an $\mathcal{E}$-weakly convergent subsequence (KR, Theorem 14.4). Any $\mathcal{E}$-weakly continuous linear functional $\ell$ on $\mathcal{L}$ is representable as $\ell(f) = \langle f, g \rangle$ for some $g \in \mathcal{E}$ (KR, Theorem 14.7). Similarly, $\{g_n\} \subset \mathcal{E}$ is $\mathcal{L}$-weakly convergent if $\{\langle f, g_n \rangle\}$ converges for each $f \in \mathcal{L}$. As $\phi(1 + x)$ satisfies the $\Delta_2$-condition, every $\mathcal{L}$-weakly continuous linear functional $\ell'$ on $\mathcal{E}$ is representable in the form $\ell'(g) = \langle f, g \rangle$ for $f \in \mathcal{L}$.[29]

We close this section by noting three useful results, the first of which is from Komunjer and Ragusa (2016). Let $\mathcal{L}_+ = \{f \in \mathcal{L} : f \geq 0 \ F_*\text{-almost everywhere}\}$.

**Lemma F.1** *Suppose that Assumption Φ(i) holds. Then:*

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

*(ii) $\mathbb{E}^{F_*}[\phi(m(U))] \leq \delta$ implies $\|m\|_\phi \leq 2 + \phi(2) + \delta$;*

*(iii) $\mathbb{E}^{F_*}[\phi(m(U))] < \infty$ if and only if $m \in \mathcal{L}_+$.*

**Proof of Lemma F.1.** Part (i) is stated on p. 961 of Komunjer and Ragusa (2016). Part (ii) follows by taking $c = \frac{1}{2}$ in the definition of $\|\cdot\|_\phi$. For part (iii), it suffices by part (ii) and the fact that $\phi(x) = +\infty$ for $x < 0$ to show $\mathbb{E}^{F_*}[\phi(m(U))] < \infty$ for all $m \in \mathcal{L}_+$. As $\phi$ satisfies the $\Delta_2$-condition under Assumption Φ(i), $m \in \mathcal{L}$ implies $\mathbb{E}^{F_*}[\phi(1 + c|m(U)|)] < \infty$ for all $c > 0$. As $\mathcal{L}$ contains constant functions and is closed under addition, for any $m \in \mathcal{L}_+$ we 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, by non-negativity of $\phi$, implies that $\mathbb{E}^{F_*}[\phi(m(U))1\!\!1\{m(U) \geq 1\}]$ is finite. Finiteness of the remaining term $\mathbb{E}^{F_*}[\phi(m(U))1\!\!1\{m(U) \leq 1\}]$ follows because $\max_{x\in[0,1]} \phi(x) = \phi(0) < \infty$ under Assumption Φ(i). ■

[29]As $\phi(1 + x)$ satisfies the $\Delta_2$-condition, in KR's notation $\mathcal{L} = L^\star_M = E_M$ with $M(x) = \phi(1 + x)$. Therefore, $\mathcal{L}$-weak convergence corresponds to what KR calls $E_M$-weak convergence on $\mathcal{E} = E_N \subseteq L^\star_N$ with $N(x) = \psi(x)$.

# G Supplementary Results and Proofs

## G.1 Notation

Throughout this Appendix, we let $\underline{K}_\delta(\theta;\gamma,P)$ and $\overline{K}_\delta(\theta;\gamma,P)$ denote the criterion functions (11) and (12). We use the notation $\underline{K}^\star_\delta(\theta;\gamma,P)$ and $\overline{K}^\star_\delta(\theta;\gamma,P)$ to denote their dual forms defined below in (A.2) and (A.3). Similarly, we let $\Delta(\theta;\gamma,P)$ to denote the primal form of the minimum divergence problem (15) and $\Delta^\star(\theta;\gamma,P)$ to denote its dual form in (16).

For $x \in \mathbb{R}^n$ and $A, B \subset \mathbb{R}^n$ we let $d(x,A) = \inf_{a\in A}\|x-a\|$ and $\vec{d}_H(A,B) = \sup_{a\in A} d(a,B)$. Let $B_\varepsilon$ denote a Euclidean ball centered at the origin with radius $\varepsilon$, where the dimension should be obvious from the context. Let $T \subseteq \mathbb{R}^n$ be a nonempty, closed convex cone with nonempty interior. Let $\partial A = \mathrm{cl}(A)\backslash \mathrm{int}(A)$ denote the boundary of $A \subset T$ (relative to $\mathbb{R}^n$) and $\partial_T A = \mathrm{cl}(\partial A \cap \mathrm{int}(T))$ denote the boundary of $A$ relative to $T$. For example, if $T = \mathbb{R}_+ \times \mathbb{R}$, and $A = \{(x,y) \in T : x^2 + y^2 \le 1\}$, then $\partial A = \{(x,y) \in T : x^2+y^2 = 1\} \cup \{0\} \times [-1,1]$ and $\partial_T A = \{(x,y) \in T : x^2 + y^2 = 1\}$.

## G.2 Preliminary Results on the Dual Form of the Criterion Functions

In this section we will show that the dual problems of (11) and (12) are

$$\underline{K}^\star_\delta(\theta;\gamma,P) = \sup_{\eta\ge 0,\zeta\in\mathbb{R},\lambda\in\Lambda} -\mathbb{E}^{F_*}\big[(\eta\phi)^\star(-k(U,\theta,\gamma)-\zeta-\lambda' g(U,\theta,\gamma))\big] - \eta\delta - \zeta - \lambda_{12}' P, \quad \text{(A.2)}$$

$$\overline{K}^\star_\delta(\theta;\gamma,P) = \inf_{\eta\ge 0,\zeta\in\mathbb{R},\lambda\in\Lambda} \mathbb{E}^{F_*}\left[(\eta\phi)^\star(k(U,\theta,\gamma)-\zeta-\lambda' g(U,\theta,\gamma))\right] + \eta\delta + \zeta + \lambda_{12}' P, \quad \text{(A.3)}$$

where $(\eta\phi)^\star$ is the convex conjugate of $x \mapsto \eta\cdot\phi(x)$. In particular, $(\eta\phi)^\star(x) = \eta\phi^\star(x/\eta)$ when $\eta > 0$. Let $\underline{\Xi}_\delta(\theta;\gamma,P)$ and $\overline{\Xi}_\delta(\theta;\gamma,P)$ denote the (possibly empty) set of solutions to the dual problems (A.2) and (A.3), respectively. The main result we prove in this subsection is the following:

**Proposition G.1** *Suppose that Assumption $\Phi$ holds. Then:*

*(i)* $\underline{K}_\delta(\theta;\gamma,P) = \underline{K}^\star_\delta(\theta;\gamma,P)$ *and* $\overline{K}_\delta(\theta;\gamma,P) = \overline{K}^\star_\delta(\theta;\gamma,P)$ *(i.e., strong duality holds);*

*(ii) Optimizing over* $(\eta,\zeta,\lambda) \in (0,\infty)\times\mathbb{R}\times\Lambda$ *yields the same value for* $\underline{K}^\star_\delta(\theta;\gamma,P)$ *and* $\overline{K}^\star_\delta(\theta;\gamma,P)$ *as optimizing over* $(\eta,\zeta,\lambda) \in \mathbb{R}_+\times\mathbb{R}\times\Lambda$*;*

*(iii) If Condition S holds at* $(\theta,\gamma,P)$ *and there is* $F$ *with* $D_\phi(F\|F_*) < \delta$ *such that (1) holds under* $F$ *at* $(\theta,\gamma,P)$*, then* $\underline{\Xi}_\delta(\theta;\gamma,P)$ *and* $\overline{\Xi}_\delta(\theta;\gamma,P)$ *are nonempty and convex;*

*(iv) If Condition S is replaced by Condition S' in (iii), then* $\underline{\Xi}_\delta(\theta;\gamma,P)$ *and* $\overline{\Xi}_\delta(\theta;\gamma,P)$ *are also compact.*

We first present some preliminary results used to derive the dual problems and verify the constraint qualification conditions. We derive the dual of (11); the derivation of the dual of (12) 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 $(\theta,\gamma)$ to simplify notation.

Consider the primal problem

$$\min_F \mathbb{E}^F[k(U)] \quad \text{subject to} \quad D_\phi(F\|F_*) \le \delta\,,\ \mathbb{E}^F[g_1(U)] \le P_1\,,\ \ldots,\ \mathbb{E}^F[g_4(U)] = 0\,. \tag{A.4}$$

The *value* of problem (A.4) is obtained by replacing the $\min$ with an $\inf$ the above display. The value is $+\infty$ if (A.4) has no solution. The criterion function $\underline{K}_\delta(\theta;\gamma,P)$ in (11) is the value of problem (A.4).

We apply duality theory as exposited in Bonnans and Shapiro (2000, Chapter 2.5). Identify each $F \in \mathcal{N}_\infty$ with $m = \mathrm{d}F/\mathrm{d}F_* \in \mathcal{L}$ (see Appendix F). Pair $\mathcal{L}$ with $\mathcal{E}$ under $\langle\cdot,\cdot\rangle$, as described in Appendix F. Define $\varphi : \mathcal{L} \times \mathbb{R}^{d+2} \to \mathbb{R} \cup \{+\infty\}$ 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 $y = (y_1,y_2,y_3) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^d$, and where $\vec{P} = (P, 0_{d_3+d_4})$, $Q_\phi(m) = \mathbb{E}^{F_*}[\phi(m(U))]$, $\langle m,1\rangle = \mathbb{E}^{F_*}[m(U)]$, $\langle m,k\rangle = \mathbb{E}^{F_*}[m(U)k(U)]$, $\langle m,g\rangle = \mathbb{E}^{F_*}[m(U)g(U)]$, and $\mathbb{I}_C : \mathbb{R}^{d+2} \to \mathbb{R}\cup\{+\infty\}$ is given by

$$\mathbb{I}_C(a_1,a_2,a_3) = \left[\begin{array}{ll} 0 & \text{if } a_1 \le 0,\ a_2 = 0, \text{ and } a_3 \in \mathbb{R}_-^{d_1} \times \{0_{d_2}\} \times \mathbb{R}_-^{d_3} \times \{0_{d_4}\}, \\ +\infty & \text{otherwise,} \end{array}\right.$$

with $\mathbb{R}_- = (-\infty, 0]$. The *primal* problem for $y \in \mathbb{R}^{d+2}$ is

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

and its value is

$$v(y) = \inf_{m\in\mathcal{L}} \varphi(m,y)\,.$$

In particular, $v(0) = \underline{K}_\delta(\theta;\gamma,P)$.

We first establish some facts about $\varphi$ and $v$. A convex function $f : X \to \mathbb{R}\cup\{+\infty\}$ is *proper* if $f(x) > -\infty$ for all $x \in X$ and $f(x) < +\infty$ for some $x \in X$.

**Lemma G.1** *Suppose that Assumption $\Phi$ holds. Then $\varphi$ is proper and convex.*

**Proof of Lemma G.1.** First note that $|\langle m,k\rangle| < +\infty$ for any $m \in \mathcal{L}$ by Hölder's inequality (see (A.1)) and Assumption $\Phi$(ii). It follows that $\varphi(m,y) > -\infty$ for all $m \in \mathcal{L}$ and $y \in \mathbb{R}^{d+2}$. Take any $m \in \mathcal{L}_+$. Then $Q_\phi(m) < +\infty$ by Lemma F.1(iii). Setting $y_1 = \delta - Q_\phi(m)$, $y_2 = 1 - \langle m,1\rangle$ and $y_3 = \vec{P} - \langle m,g\rangle$ ensures $\mathbb{I}_C(Q_\phi(m) - \delta + y_1, \langle m,1\rangle - 1 + y_2, \langle m,g\rangle - \vec{P} + y_3) = 0$, hence $\varphi(m,y) < +\infty$. Therefore, $\varphi$ is proper. Convexity of $\varphi$ follows from convexity of $m \mapsto Q_\phi(m)$ and convexity of $\mathrm{dom}\,\mathbb{I}_C \equiv \mathbb{R}_- \times \{0\} \times \mathbb{R}_-^{d_1} \times \{0_{d_2}\} \times \mathbb{R}_-^{d_3} \times \{0_{d_4}\}$. ■

Recall $\mathcal{C} = \mathbb{R}_+^{d_1} \times \{0_{d_2}\} \times \mathbb{R}_+^{d_3} \times \{0_{d_4}\}$. For $A, B \subset \mathbb{R}^n$, let $A - B = \{a - b : a \in A, b \in B\}$. Let

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

The *effective domain* of $f$ is $\operatorname{dom} f = \{x \in X : f(x) < +\infty\}$.

**Lemma G.2** *Suppose that Assumption* $\Phi$ *holds. Then:*

*(i)* $v$ *is proper, convex, and l.s.c. on* $\mathbb{R}^{d+2}$ *with* $\operatorname{dom} v = \mathcal{Y}$*;*

*(ii) A solution to the primal problem* ($\mathrm{P}_y$) *exists for each* $y \in \mathcal{Y}$*.*

**Proof of Lemma G.2.** Convexity of $v$ follows from Lemma G.1 and Bonnans and Shapiro (2000, Proposition 2.143). Note that $\mathcal{Y}$ is the set of all $y = (y_1, y_2, y_3) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ for which there exists $m_y \in \mathcal{L}$ such that

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

Take $y \in \mathcal{Y}$ and any such $m_y \in \mathcal{L}$. Then $\varphi(m_y, y) = \langle m_y, k\rangle < +\infty$ by the proof of Lemma G.1, and so $v(y) < +\infty$. Conversely, if $y \notin \mathcal{Y}$ then $\varphi(m, y) = +\infty$ for all $m \in \mathcal{L}$ and so $v(y) = +\infty$. Therefore, $\operatorname{dom} v = \mathcal{Y}$.

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

$$|v(y)| \le \sup\{|\langle k, m\rangle| : m \in \mathcal{L}, Q_\phi(m) \le \delta - y_1\} \le \|k\|_\psi(2 + \phi(2) + \delta - y_1) < \infty\,,$$

where the first inequality is by definition of $v(y)$ and the second is by inequality (A.1) and Lemma F.1(ii).

Before proving l.s.c. we first prove assertion (ii). Take any $y \in \mathcal{Y}$. Choose $\{m_n\} \subset \mathcal{L}$ such that $\varphi(m_n, y) \downarrow v(y)$ as $n \to \infty$. As $Q_\phi(m_n) \le \delta - y_1$ holds for each $n$, $\{m_n\}$ is $\|\cdot\|_\phi$-norm bounded (by Lemma F.1(ii)) and therefore has a $\mathcal{E}$-weakly convergent subsequence $\{m_{n_l}\}$ (see Appendix F). Let $m_0 \in \mathcal{L}$ denote the $\mathcal{E}$-weak limit. Under Assumption $\Phi$, 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 $\mathcal{E}$-weak convergence. We also have $\delta - y_1 \ge \liminf_{l\to\infty} Q(m_{n_l}) \ge Q(m_0)$ by Lemma F.1(i). It follows by closedness of $\operatorname{dom}\mathbb{I}_C$ that

$$\mathbb{I}_C\left(Q_\phi(m_0) - \delta + y_1, \langle m_0, 1\rangle - 1 + y_2, \langle m_0, g\rangle - \vec{P} + y_3\right) = 0\,. \tag{A.5}$$

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 ($\mathrm{P}_y$).

To prove l.s.c., we first show that $\mathcal{Y}$ is closed. Take $y \in \operatorname{cl}(\mathcal{Y})$. Choose any sequence $\{y_n\} \subset \mathcal{Y}$ converging to $y$. Note $\{v(y_n)\}$ is bounded by the argument used above to establish properness. Choose a subsequence $\{n_l\}$ for which $\lim_{l\to\infty} v(y_{n_l}) = \liminf_{n\to\infty} v(y_n)$. By part (ii), there exists a solution $m_{n_l}$ to the primal problem for each $y_{n_l}$. The sequence $\{m_{n_l}\}$ is $\|\cdot\|_\phi$-norm bounded (by Lemma F.1(ii)) and hence, taking a further subsequence if necessary, has an $\mathcal{E}$-weak limit $m_0 \in \mathcal{L}$. By similar arguments to the above, we may deduce that (A.5) holds for this $m_0$ at $y$. Therefore, $y \in \mathcal{Y}$ (establishing closedness) and $v(y) < \infty$.

To complete the proof of l.s.c., take any $y \in \mathcal{Y}$. Choose any sequence $\{y_n\} \subset \mathcal{Y}$ converging to $y$ (it is without loss of generality to consider sequences of elements of $\mathcal{Y}$, since $v(\tilde{y}) = +\infty$ for $\tilde{y} \notin \mathcal{Y}$). By the argument just used to establish closedness, we may extract a subsequence $\{n_l\}$ for which $\lim_{l\to\infty} v(y_{n_l}) = \liminf_{n\to\infty} v(y_n)$. Let $m_{n_l}$ solve the primal problem along this subsequence. Taking a further subsequence if necessary, $\{m_{n_l}\}$ has a $\mathcal{E}$-weak limit $m_0 \in \mathcal{L}$. By similar arguments to the above, we may deduce that (A.5) holds for this $m_0$ at $y$. Therefore, by $\mathcal{E}$-weak convergence we have

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

establishing l.s.c. of $v$ at any $y \in \mathrm{cl}(\mathcal{Y}) \equiv \mathcal{Y}$. Finally, l.s.c. of $v$ on $\mathbb{R}^{d+2}$ now follows from l.s.c. of $v$ at any $y \in \mathcal{Y}$ and the fact that $\mathcal{Y} \equiv \mathrm{dom}\, v$ is closed. ■

The *dual* problem of ($\mathrm{P}_y$) is (Bonnans and Shapiro, 2000, p. 96)

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

where $y^\star = (y_1^\star, y_2^\star, y_3^\star) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ and $\varphi^\star : \mathcal{E} \times \mathbb{R}^{d+2} \to \mathbb{R} \cup \{+\infty\}$ is the conjugate of $\varphi$:

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

By direct calculation,

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

where $C^o = \mathbb{R}_+ \times \mathbb{R} \times \Lambda$ is the polar cone of $C$ and $\mathbb{I}_{C^o}(y^\star) = 0$ if $y^\star \in C^o$ and $+\infty$ otherwise. Write any $y^\star \in C^o$ as $y^\star = (\eta, \zeta, \lambda) \in \mathbb{R}_+ \times \mathbb{R} \times \Lambda$. We then have

$$\begin{aligned}\varphi^\star(0, (\eta, \zeta, \lambda)) &= \sup_{m\in\mathcal{L}} \left(-\eta Q_\phi(m) - \zeta\langle m, 1\rangle - \lambda'\langle m, g\rangle - \langle m, k\rangle\right) + \eta\delta + \zeta + \lambda'\vec{P}\\ &= \sup_{m\in\mathcal{L}} \mathbb{E}^{F_*}\left[m(U)(-k(U) - \zeta - \lambda' g(U)) - \eta\phi(m(U))\right] + \eta\delta + \zeta + \lambda'\vec{P}.\end{aligned} \tag{A.6}$$

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

$$\varphi^\star(0, (\eta, \zeta, \lambda)) = \mathbb{E}^{F_*}\left[(\eta\phi)^\star(-k(U) - \zeta - \lambda' g(U))\right] + \eta\delta + \zeta + \lambda'\vec{P}, \quad (\eta, \zeta, \lambda) \in C^o,$$

where

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

As $\langle y, y^\star\rangle - \varphi^\star(0, y^\star) = -\infty$ whenever $y^\star \notin C^o$, problem ($\text{D}_y$) can therefore be expressed as

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

The *value* of this dual problem is obtained by replacing the $\max$ with $\sup$ and corresponds to the biconjugate $v^{\star\star}(y)$ of $v(y)$. Hence, $v(0) = \underline{K}_\delta(\theta;\gamma,P)$ and $v^{\star\star}(0) = \underline{K}_\delta^\star(\theta;\gamma,P)$.

**Lemma G.3** *Suppose that Assumption $\Phi$ holds. Then:*

*(i)* $\underline{K}_\delta(\theta;\gamma,P) = \underline{K}_\delta^\star(\theta;\gamma,P)$*;*

*(ii) If $0 \in \text{ri}(\mathcal{Y})$, then the set of dual solutions $\underline{\Xi}_\delta(\theta;\gamma,P)$ is nonempty and convex;*

*(iii) If $0 \in \text{int}(\mathcal{Y})$, then $\underline{\Xi}_\delta(\theta;\gamma,P)$ is also compact.*

**Proof of Lemma G.3.** For part (i), we need to show that $v(0) = v^{\star\star}(0)$. As $\varphi$ is convex by Lemma G.1 and $v$ is proper by Lemma G.2, it follows by Bonnans and Shapiro (2000, Theorem 2.144) that $v^{\star\star}(0) = \min\{v(0), \liminf_{y\to 0} v(y)\}$. But $\liminf_{y\to 0} v(y) \geq v(0)$ by Lemma G.2, and so $v(0) = v^{\star\star}(0)$, as required.

For part (ii), non-emptiness of the set of dual solutions follows by Bonnans and Shapiro (2000, Propositions 2.147 and 2.148(iii)), noting that $v$ is convex by Lemma G.2 and $v(0)$ is finite because $v$ is proper by Lemma G.2 and $0 \in \text{dom}\, v \equiv \mathcal{Y}$ by assumption. Convexity of the set of dual solutions follows by noting that, in view of ($\text{D}_y$) and (A.6), the dual objective is the pointwise infimum of affine functions of $(\eta,\zeta,\lambda)$, and is therefore concave and u.s.c. Finally, part (iii) follows from Bonnans and Shapiro (2000, Theorem 2.151 and Proposition 2.152). ■

Recall Condition S and the set $\mathcal{C}$ from Section 2.5 and Condition S' from Section 6. Define

$$\begin{aligned} \mathcal{Y}_1 &= \Big\{\vec{P} - \langle m, g\rangle : m \in \mathcal{L}_+, \langle m, 1\rangle = 1\Big\} - \mathcal{C}, \\ \mathcal{Y}_2 &= \Big\{\Big(1 - \langle m, 1\rangle, \vec{P} - \langle m, g\rangle\Big) : m \in \mathcal{L}_+\Big\} - (\{0\}\times\mathcal{C}). \end{aligned} \tag{A.7}$$

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

**Lemma G.4** *Suppose that Assumption $\Phi$ holds and Condition S holds at $(\theta,\gamma,P)$. Then:*

*(i)* $0 \in \text{ri}(\mathcal{Y}_1)$*;*

*(ii)* $0 \in \text{ri}(\mathcal{Y}_2)$*;*

*(iii) if there exists $F$ with $D_\phi(F\|F_*) < \delta$ s.t. the conditions in (A.4) hold at $\theta$, then $0 \in \text{ri}(\mathcal{Y})$.*

*Moreover, if Condition S' holds at $(\theta,\gamma,P)$ then "relative interior" can be replaced with "interior" in parts (i)–(iii).*

**Proof of Lemma G.4.** In view of Lemma F.1(iii), identify $F \in \mathcal{N}_\infty$ with its Radon–Nikodym derivative $m = \mathrm{d}F/\mathrm{d}F_* \in \mathcal{L}$. Let $\mathcal{V}_1 = \{\langle m, g\rangle : m \in \mathcal{L}_+\}$. Then

$$\vec{P} \in \mathrm{ri}(\{\mathbb{E}^F[g(U)] : F \in \mathcal{N}_\infty\} + \mathcal{C}) \iff \vec{P} \in \mathrm{ri}(\mathcal{V}_1 + \mathcal{C}) \iff 0 \in \mathrm{ri}(\mathcal{Y}_1)\,,$$

proving part (i).

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

$$\mathcal{V}_2 = \{(\langle m, 1\rangle\,, \langle m, g\rangle) : m \in \mathcal{L}_+\} + (\{0\} \times \mathcal{C}) = \mathrm{cone}\,(\{1\} \times (\mathcal{V}_1 + \mathcal{C}))\,,$$

with $\mathrm{cone}(A) = \{ta : a \in A, t \geq 0\}$. By Rockafellar (1970, Corollary 6.8.1), we have $\mathrm{ri}(\mathcal{V}_2) \supset \{1\} \times \mathrm{ri}(\mathcal{V}_1 + \mathcal{C})$. The result follows because $\vec{P} \in \mathrm{ri}(\mathcal{V}_1 + \mathcal{C})$ under Condition S.

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

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

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

$$s \begin{pmatrix} 1 \\ \vec{P} \end{pmatrix} + (1-s)v_2 = \begin{pmatrix} \langle m_v, 1\rangle \\ \langle m_v, g\rangle \end{pmatrix} + \begin{pmatrix} 0 \\ c_3 \end{pmatrix}. \tag{A.8}$$

By assumption, there exists $F \in \mathcal{N}_\delta$ with $D_\phi(F\|F_*) < \delta$ such that the moment conditions in (A.4) hold at $\theta$. 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

$$\begin{pmatrix} \langle m_v, 1\rangle \\ \langle m_v, g\rangle \end{pmatrix} = \frac{1}{\tau} \begin{pmatrix} \langle m_\tau, 1\rangle \\ \langle m_\tau, g\rangle \end{pmatrix} - \frac{1-\tau}{\tau} \begin{pmatrix} 1 \\ \vec{P} - \tilde{c} \end{pmatrix}. \tag{A.9}$$

Substituting (A.9) into (A.8) yields

$$(1 + \tau(s-1)) \begin{pmatrix} 1 \\ \vec{P} \end{pmatrix} - \tau(s-1)v_2 = \begin{pmatrix} \langle m_\tau, 1\rangle \\ \langle m_\tau, g\rangle \end{pmatrix} + \begin{pmatrix} 0 \\ \tau c_3 + (1-\tau)\tilde{c} \end{pmatrix}.$$

Note that $Q_\phi(m_\tau)$ can be made arbitrarily close to $Q_\phi(\tilde{m}) < \delta$ by choosing $\tau$ arbitrarily small. Setting

$t = 1 + \tau(s-1)$ with $\tau$ 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 $\varepsilon$-ball with center $c_0$ for some $c_0 \in \mathcal{C}$ and $\varepsilon > 0$. Then for any unit vector $x$, we have $\zeta = \lambda' c_0 + \varepsilon\lambda' x$, a contradiction. Thus, part (ii) must hold with "interior" when Condition S' holds. For part (iii), note that $\mathcal{Y} \supseteq (-\infty, \delta] \times \mathcal{Y}_2$. 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. ■

**Proof of Proposition G.1.** 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.

Part (i) follows by Lemma G.3(i). For part (ii), in view of ($\mathrm{D}_y$) and (A.6), the dual objective function, say $\ell(\eta, \zeta, \lambda)$, is the pointwise infimum of affine functions of $(\eta, \zeta, \lambda)$, and is therefore concave and u.s.c. If $\ell(0, \zeta, \lambda) = -\infty$ for all $\zeta \in \mathbb{R}$ and $\lambda \in \Lambda$, then restricting $(\eta, \zeta, \lambda)$ to $(0, \infty) \times \mathbb{R} \times \Lambda$ will not affect the dual value. Now suppose $\ell(0, \zeta, \lambda) > -\infty$ for some $(\zeta, \lambda) \in \mathbb{R} \times \Lambda$. By u.s.c. of $\ell(\cdot, \zeta, \lambda)$, we have $\ell(0, \zeta, \lambda) \geq \lim_{\eta\downarrow 0} \ell(\eta, \zeta, \lambda)$. To prove the reverse inequality, note by concavity of $\ell$ that for any $\tau \in [0, 1]$ and $\bar{\eta} > 0$, we have

$$+\infty > \ell(\tau\bar{\eta}, \zeta, \lambda) \geq (1-\tau)\ell(0, \zeta, \lambda) + \tau\ell(\bar{\eta}, \zeta, \lambda) > -\infty\,,$$

because $|\ell(\eta, \zeta, \lambda)| < +\infty$ on $(0, \infty) \times \mathbb{R} \times \Lambda$. Therefore $\lim_{\eta\downarrow 0} \ell(\eta, \zeta, \lambda) \geq \ell(0, \zeta, \lambda)$.

For part (iii), if Condition S holds and there exists $F$ with $D_\phi(F\|F_*) < \delta$ under which (1) holds, then $0 \in \mathrm{ri}(\mathcal{Y})$ by Lemma G.4. Existence and convexity of the set of dual solutions follows by Lemma G.3(ii). Finally, compactness of $\underline{\Xi}_\delta(\theta; \gamma, P)$ under Condition S' follows similarly by Lemmas G.3(iii) and Lemma G.4, proving part (iv). ■

## G.3 Additional Details on the Minimum-Divergence Problem

Recall that $\Delta(\theta; \gamma, P)$ denotes the value of the primal form of the minimum divergence problem (15) and $\Delta^\star(\theta; \gamma, P)$ denotes its dual value in (16). The main result we prove in this subsection is the following:

**Proposition G.2** *Suppose that Assumption $\Phi$ holds. Then:*

*(i)* $\Delta(\theta;\gamma,P) = \Delta^\star(\theta;\gamma,P)$ *(i.e., strong duality holds);*

*(ii) If Condition S holds at $(\theta,\gamma,P)$, then the set of dual solutions is nonempty and convex;*

*(iii) If Condition S' holds at $(\theta,\gamma,P)$, then the set of dual solutions is compact.*

We first present some preliminary results. We again drop dependence of $k$ and $g$ on $(\theta,\gamma)$ to simplify notation. Define $\varphi_2 : \mathcal{L} \times \mathbb{R}^{d+1} \to \mathbb{R} \cup \{+\infty\}$ by

$$\varphi_2(m,y) = Q_\phi(m) + \mathbb{I}_{C_2}\left(\langle m,1\rangle - 1 + y_1, \langle m,g\rangle - \vec{P} + y_2\right),$$

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

$$\mathbb{I}_{C_2}(y_1,y_2) = \left[\begin{array}{ll} 0 & \text{if } y_1 = 0, \text{ and } y_2 \in \mathbb{R}^{d_1}_- \times \{0_{d_2}\} \times \mathbb{R}^{d_3}_- \times \{0_{d_4}\}, \\ +\infty & \text{otherwise.} \end{array}\right.$$

The primal problem for $y \in \mathbb{R}^{d+1}$ is

$$\min_{m \in \mathcal{L}} \varphi_2(m,y) \tag{P$'_y$}$$

and its value is

$$v_2(y) = \inf_{m \in \mathcal{L}} \varphi_2(m,y)\,.$$

In particular, $v_2(0) = \Delta(\theta,\gamma,P)$. Recall the set $\mathcal{Y}_2$ defined in (A.7)

**Lemma G.5** *Suppose that Assumption $\Phi$ holds. Then:*

*(i) $\varphi_2$ is proper and convex;*

*(ii) $v_2$ is proper, convex, and* $\operatorname{dom} v_2 = \mathcal{Y}_2$*;*

*(iii) A solution to the primal problem (P$'_y$) exists and is unique for each $y \in \mathcal{Y}_2$.*

**Proof of Lemma G.5.** Properness and convexity of $\varphi_2$ and $v_2$ follow by similar arguments to Lemma G.1 and G.2, and $\operatorname{dom} v_2 = \mathcal{Y}_2$ follows by similar arguments to Lemma G.2.

For part (iii), take any $y \in \mathcal{Y}_2$. Choose $\{m_n\} \subset \mathcal{L}$ such that $\varphi_2(m_n,y) \downarrow v_2(y)$ as $n \to \infty$. Without loss of generality we may assume $\varphi_2(m_n,y) < \infty$ for each $n$. Then as $\varphi_2(m_n,y) = Q_\phi(m_n)$, the sequence $\{m_n\}$ is $\|\cdot\|_\phi$-norm bounded (by Lemma F.1(ii)) and therefore has a $\mathcal{E}$-weakly convergent subsequence $\{m_{n_l}\}$ (see Appendix F). Let $m_0 \in \mathcal{L}$ denote the $\mathcal{E}$-weak limit. 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 $\mathcal{E}$-weak convergence. Therefore,

$$\mathbb{I}_{C_2}\left(\langle m_0,1\rangle - 1 + y_1, \langle m_0,g\rangle - \vec{P} + y_2\right) = 0 \tag{A.10}$$

and so $\varphi_2(m_0,y) = Q_\phi(m_0)$. It follows by Lemma F.1(i) that $v_2(y) = \lim_{l\to\infty} Q_\phi(m_{n_l}) \geq Q_\phi(m_0) \geq v_2(y)$, where the final inequality is by definition of $v_2(y)$. Uniqueness of the primal solution follows by strict convexity of $\phi$. ■

By similar arguments to Appendix G.2, the dual problem of ($P'_y$) is

$$\max_{\zeta \in \mathbb{R}, \lambda \in \Lambda} \zeta y_1 + \lambda' y_2 - \mathbb{E}^{F_*}\left[\phi^\star(-\zeta - \lambda' g(U))\right] - \zeta - \lambda'_{12} P \,. \tag{$D'_y$}$$

The *value* of ($D'_y$) is obtained by replacing the max with a sup and corresponds to the biconjugate $v_2^{\star\star}(y)$ of $v_2(y)$. Hence, $v_2(0) = \Delta(\theta;\gamma,P)$ and $v_2^{\star\star}(0) = \Delta^\star(\theta;\gamma,P)$.

**Proof of Proposition G.2.** For part (i), we need to show $v_2(0) = v_2^{\star\star}(0)$. As $\varphi_2$ is convex by Lemma G.5 and $v_2(y) \geq 0$ for all $y \in \mathbb{R}^{d+1}$, it follows by Bonnans and Shapiro (2000, Theorem 2.144) that $v_2^{\star\star}(0) = \min\{v_2(0), \liminf_{y \to 0} v_2(y)\}$. Therefore, it suffices to show that $v_2(y)$ is l.s.c. at $y = 0$.

To do so, let $\{y_n\}$ be any sequence converging to 0. It is without loss of generality to assume $\liminf_{n\to\infty} v_2(y_n) < +\infty$ (otherwise there is nothing to prove). Extract a subsequence $\{n_l\}$ along which $v_2(y_{n_l}) < +\infty$ and $\lim_{l\to\infty} v_2(y_{n_l}) = \liminf_{n\to\infty} v_2(y_n)$. The primal problem at $y_{n_l}$ has a solution $m_{n_l}$ for each $l$ by Lemma G.5. As $v_2(y_{n_l}) = Q_\phi(m_{n_l}) \leq C$ for all $l$ and a finite positive constant $C$, the sequence $\{m_{n_l}\}$ is $\|\cdot\|_\phi$-norm bounded (by Lemma F.1(ii)). Extracting a further subsequence if necessary, we may assume the sequence $\{m_{n_l}\}$ is $\mathcal{E}$-weakly convergent to some $m_0 \in \mathcal{L}$. By similar arguments to the proof of Lemma G.5, we may deduce that (A.10) holds for this $m_0$ at $y = 0$. Then by Lemma F.1(i), we have

$$\liminf_{n\to\infty} v_2(y_n) = \lim_{l\to\infty} v_2(y_{n_l}) = \lim_{l\to\infty} Q_\phi(m_{n_l}) \geq Q_\phi(m_0) \geq v_2(0)\,,$$

as required. Note by way of contradiction that the preceding argument also implies that $\liminf_{n\to\infty} v_2(y_n) = +\infty$ for all sequences $y_n \to 0$ whenever $v_2(0) = +\infty$.

For parts (ii) and (iii), Lemma G.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. Parts (ii) and (iii) now follow by similar arguments to Lemma G.3(ii)(iii). ■

## G.4 Stability of Constraint Qualifications under Perturbations

**Lemma G.6** *Suppose that Assumption $\Phi$ holds and Condition S' holds at $(\theta,\gamma,P)$. 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 G.6.** Recall $\vec{P} = (P, 0_{d_3+d_4})$ and let $\vec{\tilde{P}} = (\tilde{P}, 0_{d_3+d_4})$. By Condition S', there exists $\varepsilon > 0$ such that $B_{2\varepsilon} \subseteq (\{\mathbb{E}^F[g(U,\theta,\gamma)] - \vec{P} : F \in \mathcal{N}_\infty\} + \mathcal{C})$. Then for any $\tilde{P}$ for which $\|P - \tilde{P}\| < \varepsilon$, we have $\|(\mathbb{E}^F[g(U,\theta,\gamma)] - \vec{P}) - (\mathbb{E}^F[g(U,\theta,\gamma)] - \vec{\tilde{P}})\| < \varepsilon$ for all $F \in \mathcal{N}_\infty$. Hence, $B_\varepsilon \subseteq (\{\mathbb{E}^F[g(U,\theta,\gamma)] - \vec{\tilde{P}} : F \in \mathcal{N}_\infty\} + \mathcal{C})$. ■

**Lemma G.7** *Suppose that Assumption $\Phi$ holds, each entry of $g$ is $\mathcal{E}$-continuous in $(\theta,\gamma)$, and Condition S' holds at $(\theta,\gamma,P)$. Then there exists a neighborhood $N$ of $(\theta,\gamma,P)$ 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 G.7.** Recall $\vec{P} = (P, 0_{d_3+d_4})$. As Condition S' holds at $(\theta, \gamma, P)$, there exists sufficiently large $\delta$ such that $0 \in \text{int}(\{\mathbb{E}^F[g(U,\theta,\gamma)] - \vec{P} : F \in \mathcal{N}_\delta\} + \mathcal{C})$. (To see this, take a sufficiently small hypercube $\mathcal{H}$ with $0 \in \text{int}(\mathcal{H})$ and $\mathcal{H} \subseteq \text{int}(\{\mathbb{E}^F[g(U,\theta,\gamma)] - \vec{P} : F \in \mathcal{N}_\infty\} + \mathcal{C})$, identify a density $F \in \mathcal{N}_\infty$ with each vertex of $\mathcal{H}$, and take $\delta$ to be the largest $\phi$-divergence from $F_*$ of each of the densities at the vertex.) Therefore, we may choose $\varepsilon > 0$ such that $B_{4\varepsilon} \subseteq \text{int}(\{\mathbb{E}^F[g(U,\theta,\gamma)] - \vec{P} : F \in \mathcal{N}_\delta\} + \mathcal{C})$.

Let $\mathcal{M}_\delta = \{\frac{\mathrm{d}F}{\mathrm{d}F_*} : F \in \mathcal{N}_\delta\}$. Then $\mathcal{M}_\delta$ is a $\|\cdot\|_\phi$-bounded subset of $\mathcal{L}$ by Lemma F.1(ii). By $\mathcal{E}$-continuity, there exists a neighborhood $N_1$ of $(\theta,\gamma)$ such that for any $(\tilde{\theta},\tilde{\gamma}) \in N_1$ and with $r$ denoting any entry of $g_1,\dots,g_4$, we have

$$\|r(\cdot,\theta,\gamma) - r(\cdot,\tilde{\theta},\tilde{\gamma})\|_\psi < \frac{\varepsilon}{\sqrt{d}(2+\phi(2)+\delta)} \,.$$

It follows by inequality (A.1) and Lemma F.1(ii) that for any $(\tilde{\theta},\tilde{\gamma}) \in N_1$, we have

$$\sup_{m \in \mathcal{M}_\delta} |\mathbb{E}^{F_*}[m(U)r(U,\theta,\gamma)] - \mathbb{E}^{F_*}[m(U)r(U,\tilde{\theta},\tilde{\gamma})]| \le \frac{\varepsilon}{\sqrt{d}}.$$

Let $N_2$ be an $\varepsilon$-neighborhood of $P$. Then for any $F \in \mathcal{N}_\delta$ and any $(\tilde{\theta},\tilde{\gamma},\tilde{P}) \in N_1 \times N_2$, we have

$$\|(\mathbb{E}^F[m(U)g(U,\theta,\gamma)] - \vec{P}) - (\mathbb{E}^F[m(U)g(U,\tilde{\theta},\tilde{\gamma})] - \vec{\tilde{P}})\| < 2\varepsilon \,,$$

with $\vec{\tilde{P}} = (\tilde{P}, 0_{d_3+d_4})$, and so

$$B_{2\varepsilon} \subseteq \text{int}(\{\mathbb{E}^F[g(U,\tilde{\theta},\tilde{\gamma})] - \vec{\tilde{P}} : F \in \mathcal{N}_\delta\} + \mathcal{C}) \subseteq \text{int}(\{\mathbb{E}^F[g(U,\tilde{\theta},\tilde{\gamma})] - \vec{\tilde{P}} : F \in \mathcal{N}_\infty\} + \mathcal{C}) \,,$$

as required. ■

## G.5 Continuity of the Optimal Values

**Lemma G.8** *Suppose that Assumption $\Phi$ holds and $\mathbb{E}^{F_*}[\phi^\star(a_1 + a_2' g(U,\theta,\gamma))]$ is continuous in $(\theta,\gamma)$ for every $(a_1,a_2) \in \mathbb{R}\times\mathbb{R}^d$. Then $\Delta(\tilde{\theta};\tilde{\gamma},\tilde{P})$ and $\Delta^\star(\tilde{\theta};\tilde{\gamma},\tilde{P})$ are continuous in $(\tilde{\theta},\tilde{\gamma},\tilde{P})$ at any point $(\theta,\gamma,P)$ at which Condition S' holds.*

**Proof of Lemma G.8.** In view of Proposition G.2(i), it suffices to establish continuity of $\Delta^\star(\tilde{\theta};\tilde{\gamma},\tilde{P})$.

Fix some $(\theta,\gamma,P)$ at which Conditions S' holds. Then by definition of Condition S', we have $\Delta(\theta;\gamma,P) < \infty$ and hence $\Delta^\star(\theta;\gamma,P) < \infty$. To simplify notation, let $\xi = (\zeta,\lambda)$ and $\Xi = \mathbb{R}\times\Lambda$ for the remainder of this proof. Note

$$\xi \mapsto L(\xi;\theta,\gamma,P) := -\mathbb{E}^{F_*}[\phi^\star(-\zeta - \lambda' g(U,\theta,\gamma))] - \zeta - \lambda_{12}' P$$

is the pointwise infimum of affine functions of $\xi$ and is therefore concave and u.s.c. By Proposition G.2(iii), $L(\,\cdot\,;\theta,\gamma,P)$ has a nonempty, convex, and compact set of maximizers $\Xi_0 \subset \Xi$. Fix $\varepsilon > 0$ and let $\Xi_0^\varepsilon = \{\xi \in \Xi : d(\xi, \Xi_0) \le \varepsilon\}$.

By continuity of $(\theta,\gamma) \mapsto \mathbb{E}^{F_*}[\phi^\star(\zeta + \lambda' g(U,\theta,\gamma))]$, we have $L(\xi;\tilde\theta,\tilde\gamma,\tilde P) \to L(\xi;\theta,\gamma,P)$ as $(\tilde\theta,\tilde\gamma,\tilde P) \to (\theta,\gamma,P)$. By concavity of $L$ in $\xi$, convergence may be strengthened to hold uniformly over the compact set $\Xi_0^\varepsilon$ (Rockafellar, 1970, Theorem 10.8), and so

$$\sup_{\xi \in \Xi_0^\varepsilon} L(\xi;\tilde\theta,\tilde\gamma,\tilde P) \to \Delta^\star(\theta;\gamma,P) \quad \text{as} \quad (\tilde\theta,\tilde\gamma,\tilde P) \to (\theta,\gamma,P)\,. \tag{A.11}$$

By u.s.c. of $L(\,\cdot\,;\theta,\gamma,P)$, definition of $\Xi_0$, and the fact that $\Xi_0 \cap \partial_\Xi \Xi_0^\varepsilon = \emptyset$, we also have that

$$\Delta^\star(\theta;\gamma,P) > \sup_{\xi \in \partial_\Xi \Xi_0^\varepsilon} L(\xi;\theta,\gamma,P)\,.$$

It follows that there exists a neighborhood $N$ of $(\theta,\gamma,P)$ such that for any $(\tilde\theta,\tilde\gamma,\tilde P) \in N$,

$$\sup_{\xi \in \Xi_0^\varepsilon} L(\xi;\tilde\theta,\tilde\gamma,\tilde P) > \sup_{\xi \in \partial_\Xi \Xi_0^\varepsilon} L(\xi;\tilde\theta,\tilde\gamma,\tilde P)\,.$$

By standard arguments for maximizers of concave functions (e.g., the proof of Theorem 2.7 of Newey and McFadden (1994)), whenever $(\tilde\theta,\tilde\gamma,\tilde P) \in N$ we have that

$$\sup_{\xi \in \Xi_0^\varepsilon} L(\xi;\tilde\theta,\tilde\gamma,\tilde P) = \sup_{\xi \in \Xi} L(\xi;\tilde\theta,\tilde\gamma,\tilde P) \equiv \Delta^\star(\tilde\theta;\tilde\gamma,\tilde P)\,.$$

The result now follows by (A.11). ■

We now establish a similar result for the objectives (11) and (12) and their duals. To simplify notation, in the remainder of this subsection fix $(\theta,\gamma,P)$ and let $\xi = (\eta,\zeta,\lambda)$, $\Xi = \mathbb{R}_+ \times \mathbb{R} \times \Lambda$, $\underline{\Xi}_\delta = \underline{\Xi}_\delta(\theta;\gamma,P)$ and $\overline{\Xi}_\delta = \overline{\Xi}_\delta(\theta;\gamma,P)$. If $\underline{\Xi}_\delta$ and $\overline{\Xi}_\delta$ are compact, then for each $\varepsilon > 0$ we may cover them by compact sets $\underline{\Xi}_\delta^\varepsilon \subset \Xi$ and $\overline{\Xi}_\delta^\varepsilon \subset \Xi$ formed as the union of finitely many hypercubes with edges parallel to the coordinate axes, so that $d(\xi,\underline{\Xi}_\delta) \le \varepsilon$ for all $\xi \in \underline{\Xi}_\delta^\varepsilon$, $d(\xi,\overline{\Xi}_\delta) \le \varepsilon$ for all $\xi \in \overline{\Xi}_\delta^\varepsilon$, $(\partial_\Xi \underline{\Xi}_\delta^\varepsilon) \cap \underline{\Xi}_\delta = \emptyset$, and $(\partial_\Xi \overline{\Xi}_\delta^\varepsilon) \cap \overline{\Xi}_\delta = \emptyset$.

**Lemma G.9** *Suppose that Assumptions $\Phi$ and M(i),(ii) hold, Condition S' holds at $(\theta,\gamma,P)$, and $\Delta(\theta;\gamma,P) < \delta$. Then:*

*(i) $\underline{K}_\delta(\tilde\theta;\tilde\gamma,\tilde P)$, $\overline{K}_\delta(\tilde\theta;\tilde\gamma,\tilde P)$, $\underline{K}_\delta^\star(\tilde\theta;\tilde\gamma,\tilde P)$, and $\overline{K}_\delta^\star(\tilde\theta;\tilde\gamma,\tilde P)$ are continuous in $(\tilde\theta,\tilde\gamma,\tilde P)$ at $(\theta,\gamma,P)$;*

*(ii) For each $\varepsilon > 0$ there exists a neighborhood $N$ of $(\theta,\gamma,P)$ such that $\underline{\Xi}_\delta(\tilde\theta;\tilde\gamma,\tilde P) \subseteq \underline{\Xi}_\delta^\varepsilon$ and $\overline{\Xi}_\delta(\tilde\theta;\tilde\gamma,\tilde P) \subseteq \overline{\Xi}_\delta^\varepsilon$ for each $(\tilde\theta,\tilde\gamma,\tilde P) \in N$;*

*(iii) $\vec d_H(\underline{\Xi}_\delta(\tilde\theta;\tilde\gamma,\tilde P),\underline{\Xi}_\delta) \to 0$ and $\vec d_H(\overline{\Xi}_\delta(\tilde\theta;\tilde\gamma,\tilde P),\overline{\Xi}_\delta) \to 0$ as $(\tilde\theta,\tilde\gamma,\tilde P) \to (\theta,\gamma,P)$.*

**Proof of Lemma G.9.** We prove the result for $\underline{K}_\delta^\star$ and $\underline{\Xi}_\delta$; the result for $\overline{K}_\delta^\star$ and $\overline{\Xi}_\delta$ follows similarly.

Step 1: Preliminaries. In view of Proposition G.1(i), it suffices to establish continuity of $\underline{K}^\star_\delta(\theta;\gamma,P)$. Lemmas G.7 and G.8 imply there is a neighborhood $N'$ of $(\theta,\gamma,P)$ such that Condition S' holds at $(\tilde\theta,\tilde\gamma,\tilde P)$ and $\Delta(\tilde\theta,\tilde\gamma,\tilde P) < \delta$ for each $(\tilde\theta,\tilde\gamma,\tilde P) \in N'$. By Proposition G.1(iii)(iv), 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 $\Xi$ for each $(\tilde\theta,\tilde\gamma,\tilde P) \in N'$.

The dual objective function

$$\xi \mapsto L(\xi;\theta,\gamma,P) := -\mathbb{E}^{F_*}[(\eta\phi)^\star(-k(U,\theta,\gamma)-\zeta-\lambda' g(U,\theta,\gamma))] - \eta\delta - \zeta - \lambda'_{12}P$$

is the pointwise infimum of affine functions of $\xi = (\eta,\zeta,\lambda)$ and is therefore concave and u.s.c. Hence,

$$4a := \underline{K}^\star_\delta(\theta;\gamma,P) - \sup_{\xi\in\partial_\Xi \underline{\Xi}^\varepsilon_\delta} L(\xi;\theta,\gamma,P) > 0\,. \tag{A.12}$$

The remaining steps of the proof depend on whether or not $\min\{\eta : (\eta,\zeta,\lambda) \in \underline{\Xi}_\delta\} > 0$.

Step 2: Proof of parts (i) and (ii) when $\min\{\eta : (\eta,\zeta,\lambda) \in \underline{\Xi}_\delta\} > 0$. W.l.o.g. we may choose $\underline{\Xi}^\varepsilon_\delta$ so that $\min\{\eta : (\eta,\zeta,\lambda) \in \underline{\Xi}^\varepsilon_\delta\} > 0$. For any $\xi = (\eta,\zeta,\lambda)$ with $\eta > 0$,

$$L(\xi;\tilde\theta,\tilde\gamma,\tilde P) = -\eta\mathbb{E}^{F_*}\left[\phi^\star\left(\tfrac{k(U,\tilde\theta,\tilde\gamma)+\zeta+\lambda' g(U,\tilde\theta,\tilde\gamma)}{-\eta}\right)\right] - \eta\delta - \zeta - \lambda'_{12}\tilde P\,.$$

By Assumption M(ii), for any $\xi \in \Xi$ we have $L(\xi;\tilde\theta,\tilde\gamma,\tilde P) \to L(\xi;\theta,\gamma,P)$ as $(\tilde\theta,\tilde\gamma,\tilde P) \to (\theta,\gamma,P)$. By concavity of $L$ in $\xi$, convergence may be strengthened to hold uniformly over the compact set $\underline{\Xi}^\varepsilon_\delta$ (Rockafellar, 1970, Theorem 10.8) and so, in particular,

$$\sup_{\xi\in\underline{\Xi}^\varepsilon_\delta} L(\xi;\tilde\theta,\tilde\gamma,\tilde P) \to \underline{K}^\star_\delta(\theta;\gamma,P) \text{ as } (\tilde\theta,\tilde\gamma,\tilde P) \to (\theta,\gamma,P). \tag{A.13}$$

It follows that there exists a neighborhood $N''$ of $(\theta,\gamma,P)$ such that for $(\tilde\theta,\tilde\gamma,\tilde P) \in N' \cap N''$, we have

$$\sup_{\xi\in\underline{\Xi}^\varepsilon_\delta} L(\xi;\tilde\theta,\tilde\gamma,\tilde P) > \sup_{\xi\in\partial_\Xi\underline{\Xi}^\varepsilon_\delta} L(\xi;\tilde\theta,\tilde\gamma,\tilde P)\,.$$

By similar arguments to the proof of Lemma G.8 we may deduce that $\underline{\Xi}_\delta(\tilde\theta;\tilde\gamma,\tilde P) \subseteq \underline{\Xi}^\varepsilon_\delta$ holds on $N := N' \cap N''$. This proves part (ii). Continuity (part (i)) now follows by (A.13).

Step 3: Proof of part (ii) when $\min\{\eta : (\eta,\zeta,\lambda) \in \underline{\Xi}_\delta\} = 0$. Choose $\xi_0 := (0,\underline{\zeta},\underline{\lambda}) \in \underline{\Xi}_\delta$. Let $\xi_\eta = (\eta,\underline{\zeta},\underline{\lambda})$ for $\eta > 0$. As in the proof of Proposition G.1, it follows by concavity and u.s.c. of $L$ in $\xi$ that

$$\lim_{\eta\downarrow 0} L(\xi_\eta;\theta,\gamma,P) = L(\xi_0;\theta,\gamma,P) = \underline{K}^\star_\delta(\theta;\gamma,P)\,.$$

For any $\varepsilon_0 \in (0,a)$, choose $\bar\eta > 0$ such that $L(\xi_{\bar\eta};\theta,\gamma,P) > \underline{K}^\star_\delta(\theta;\gamma,P) - \varepsilon_0$ and $\xi_{\bar\eta} \in \text{int}(\underline{\Xi}^\varepsilon_\delta)$. By

Assumption M(ii), there is a neighborhood $N''$ of $(\theta, \gamma, P)$ such that for all $(\tilde{\theta}, \tilde{\gamma}, \tilde{P}) \in N''$ we have

$$L(\xi_{\bar{\eta}}; \tilde{\theta}, \tilde{\gamma}, \tilde{P}) > \underline{K}_\delta^\star(\theta; \gamma, P) - 2\varepsilon_0. \tag{A.14}$$

We now argue by contradiction that the inequality

$$\sup_{\xi \in \partial_\Xi \Xi_\delta^\varepsilon} L(\xi; \theta, \gamma, P) \geq \sup_{\xi \in \partial_\Xi \Xi_\delta^\varepsilon} L(\xi; \tilde{\theta}, \tilde{\gamma}, \tilde{P}) - 2\varepsilon_0 \tag{A.15}$$

holds for all $(\tilde{\theta}, \tilde{\gamma}, \tilde{P})$ in a neighborhood $N'''$ of $(\theta, \gamma, P)$. Suppose that there is $\varepsilon_1 > 0$ and $(\theta_n, \gamma_n, P_n) \to (\theta, \gamma, P)$ along which

$$\sup_{\xi \in \partial_\Xi \Xi_\delta^\varepsilon} L(\xi; \theta, \gamma, P) \leq \sup_{\xi \in \partial_\Xi \Xi_\delta^\varepsilon} L(\xi; \theta_n, \gamma_n, P_n) - \varepsilon_1 \,. \tag{A.16}$$

For each $n \geq 1$, choose $\xi_n \in \arg\sup_{\xi \in \partial_\Xi \Xi_\delta^\varepsilon} L(\xi; \theta_n, \gamma_n, P_n)$. As $\partial_\Xi \Xi_\delta^\varepsilon$ is compact, take a convergent subsequence $\{\xi_{n_l}\}$ and let $\xi^* = (\eta^*, \zeta^*, \lambda^*) \in \partial_\Xi \Xi_\delta^\varepsilon$ denote its limit point. Suppose $\eta^* > 0$. Then as $L(\,\cdot\,; \theta_n, \gamma_n, P_n)$ converges uniformly to $L(\,\cdot\,; \theta, \gamma, P)$ on compact subsets of $(0, \infty) \times \mathbb{R} \times \mathbb{R}^d$, we obtain $\lim_{l \to \infty} L(\xi_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \leq \sup_{\xi \in \partial_\Xi \Xi_\delta^\varepsilon} L(\xi; \theta, \gamma, P)$, which contradicts (A.16).

Conversely, if $\eta^* = 0$, define $\xi_\eta^* = (\eta, \zeta^*, \lambda^*)$. Fix any small $\varepsilon_2 > 0$ so that $\xi_{\varepsilon_2}^* \in \partial_\Xi \Xi_\delta^\varepsilon$. By u.s.c. and concavity of $L(\,\cdot\,; \theta, \gamma, P)$, we may choose $\varepsilon_2$ sufficiently small that

$$L(\xi_{\varepsilon_2}^*; \theta, \gamma, P) - L(\xi_{2\varepsilon_2}^*; \theta, \gamma, P) < \frac{1}{2}\varepsilon_1 \,. \tag{A.17}$$

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,

$$L(\varepsilon_2, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \geq \tau_{n_l} L(\xi_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) + (1 - \tau_{n_l}) L(2\varepsilon_2, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \,.$$

which rearranges to yield

$$L(\xi_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \leq \frac{1}{\tau_{n_l}} \left( L(\varepsilon_2, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) - (1 - \tau_{n_l}) L(2\varepsilon_2, \zeta_{n_l}, \lambda_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \right) \,.$$

As $L(\,\cdot\,; \theta_n, \gamma_n, P_n)$ converges uniformly to $L(\,\cdot\,; \theta, \gamma, P)$ on compact subsets of $(0, \infty) \times \mathbb{R} \times \mathbb{R}^d$, we obtain

$$\lim_{l \to \infty} L(\xi_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) \leq 2L(\xi_{\varepsilon_2}^*; \theta, \gamma, P) - L(\xi_{2\varepsilon_2}^*; \theta, \gamma, P) \,.$$

It follows by (A.17) that for all $l$ sufficiently large we have

$$L(\xi_{n_l}; \theta_{n_l}, \gamma_{n_l}, P_{n_l}) < \sup_{\xi \in \partial_\Xi \Xi_\delta^\varepsilon} L(\xi; \theta, \gamma, P) + \varepsilon_1 \,,$$

which contradicts (A.16). This completes the proof of inequality (A.15).

It now follows from displays (A.12), (A.14), and (A.15) that on $N' \cap N'' \cap N'''$ we have

$$\begin{aligned} L(\xi_{\bar\eta}; \tilde\theta, \tilde\gamma, \tilde P) > \underline{K}^\star_\delta(\theta; \gamma, P) - 2\varepsilon_0 &= \sup_{\xi \in \partial_\Xi \Xi^\varepsilon_\delta} L(\xi; \theta, \gamma, P) + 4a - 2\varepsilon_0 \\ &\geq \sup_{\xi \in \partial_\Xi \Xi^\varepsilon_\delta} L(\xi; \tilde\theta, \tilde\gamma, \tilde P) + 4(a - \varepsilon_0) \,. \end{aligned}$$

As $a - \varepsilon_0 > 0$, the inequality $\sup_{\xi \in \Xi^\varepsilon_\delta} L(\xi; \tilde\theta, \tilde\gamma, \tilde P) > \sup_{\xi \in \partial_\Xi \Xi^\varepsilon_\delta} L(\xi; \tilde\theta, \tilde\gamma, \tilde P)$ holds on $N := N' \cap N'' \cap N'''$. It now follows by standard arguments for maximizers of concave objective functions (e.g., the proof of Theorem 2.7 of Newey and McFadden (1994)) that $\Xi_\delta(\tilde\theta; \tilde\gamma, \tilde P) \subseteq \Xi^\varepsilon_\delta$ 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

$$\underline{K}^\star_\delta(\tilde\theta; \tilde\gamma, \tilde P) = \sup_{\xi \in \Xi^\varepsilon_\delta} L(\xi; \tilde\theta, \tilde\gamma, \tilde P) \tag{A.18}$$

by Step 3, so by (A.14) we have $\underline{K}^\star_\delta(\tilde\theta; \tilde\gamma, \tilde P) \geq L(\xi_{\bar\eta}; \tilde\theta, \tilde\gamma, \tilde P) > \underline{K}^\star_\delta(\theta; \gamma, P) - 2\varepsilon_0$ for $(\tilde\theta, \tilde\gamma, \tilde P) \in N$, proving l.s.c.

To establish u.s.c., for any $\varepsilon_0 > 0$ one may deduce (by similar arguments used to establish (A.15) in Step 3) that there is a neighborhood $N''''$ of $(\theta, \gamma, P)$ such that for any $(\tilde\theta, \tilde\gamma, \tilde P) \in N''''$ we have

$$\sup_{\xi \in \Xi^\varepsilon_\delta} L(\xi; \theta, \gamma, P) \geq \sup_{\xi \in \Xi^\varepsilon_\delta} L(\xi; \tilde\theta, \tilde\gamma, \tilde P) - \varepsilon_0 \tag{A.19}$$

holds. It follows by (A.18) and (A.19) that on $N \cap N''''$, we have

$$\underline{K}^\star_\delta(\tilde\theta; \tilde\gamma, \tilde P) = \sup_{\xi \in \Xi^\varepsilon_\delta} L(\xi; \tilde\theta, \tilde\gamma, \tilde P) \leq \sup_{\xi \in \Xi^\varepsilon_\delta} L(\xi; \theta, \gamma, P) + \varepsilon_0 = \underline{K}^\star_\delta(\theta; \gamma, P) + \varepsilon_0 \,.$$

Step 5: Proof of part (iii). By part (ii), for each $\varepsilon > 0$ there exists a neighborhood $N$ such that $\Xi_\delta(\tilde\theta; \tilde\gamma, \tilde P) \subseteq \Xi^\varepsilon_\delta$ for all $(\tilde\theta, \tilde\gamma, \tilde P) \in N$. Therefore, $\vec d_H(\Xi_\delta(\tilde\theta; \tilde\gamma, \tilde P), \Xi_\delta) \leq \vec d_H(\Xi^\varepsilon_\delta, \Xi_\delta) \leq \varepsilon$ for all $(\tilde\theta, \tilde\gamma, \tilde P) \in N$. ■

## G.6 Proofs for Appendix A

**Proof of Proposition A.1.** We prove only the result for $\underline{K}^\Pi_\delta$; the result for $\overline{K}^\Pi_\delta$ follows similarly.

Dropping dependence of $k$ and $g$ on $(\theta, \gamma)$ to simplify notation, we have

$$\begin{aligned} \underline{K}^\Pi_\delta(\theta; \gamma, P) &= \inf_{F \in \mathcal{N}^\Pi_\delta} \mathbb{E}^F[k^\Pi(U)] \quad \text{subject to} \quad \mathbb{E}^F[g^\Pi_1(U)] \leq P_1 \,, \ldots, \mathbb{E}^F[g^\Pi_4(U)] = 0 \\ &\geq \inf_{F \in \mathcal{N}_\delta} \mathbb{E}^F[k^\Pi(U)] \quad \text{subject to} \quad \mathbb{E}^F[g^\Pi_1(U)] \leq P_1 \,, \ldots, \mathbb{E}^F[g^\Pi_4(U)] = 0 \,, \end{aligned} \tag{A.20}$$

where the first line uses $\Pi$-invariance of $F \in \mathcal{N}_\delta^\Pi$ and the second is because $\mathcal{N}_\delta^\Pi \subseteq \mathcal{N}_\delta$. Problem (A.20) is an optimization over $\mathcal{N}_\delta$ may therefore be restated as problem (34) by virtue of Proposition 2.1. The right-hand side of (34) is therefore a lower bound for $\underline{K}_\delta^\Pi(\theta;\gamma,P)$.

To establish equality, first suppose problem (A.20) is infeasible. Then problem (33) must also be infeasible because $\mathcal{N}_\delta \supseteq \mathcal{N}_\delta^\Pi$, so the value of problem (33) is $+\infty$. By Proposition 2.1, the value of the dual program to (A.20), and hence the right-hand side of (34), must also be $+\infty$.

Now suppose that problem (A.20) is feasible. We claim that there exists a minimizing $F \in \mathcal{N}_\delta$ that is $\Pi$-invariant, so it is without loss of generality to optimize over $\mathcal{N}_\delta$ rather than $\mathcal{N}_\delta^\Pi$. First note by Lemma G.2 there exists a (not necessarily $\Pi$-invariant) minimizing $F \in \mathcal{N}_\delta$, say $F_0$. Let $m_0 = \frac{\mathrm{d}F_0}{\mathrm{d}F_*}$ and let $\kappa_0 = \mathbb{E}^{F_0}[k^\Pi(U)]$ denote the minimizing value of (A.20). Note that $\kappa_0$ must also be the value of the dual program (34) by virtue of Proposition 2.1. Define $m_\Pi(u) = \frac{1}{|\Pi|}\sum_{\varpi\in\Pi} m_0(\varpi u)$ and $F_\Pi$ by $\mathrm{d}F_\Pi = m_\Pi \mathrm{d}F_*$.

We have $\mathbb{E}^{F_*}[m_\Pi(U)] = \frac{1}{|\Pi|}\sum_{\varpi\in\Pi}\mathbb{E}^{F_*}[m_0(\varpi U)] = \frac{1}{|\Pi|}\sum_{\varpi\in\Pi}\mathbb{E}^{F_*}[m_0(U)] = 1$ by $\Pi$-invariance of $F_*$, so $F_\Pi$ is a probability measure. By convexity of $\phi$ we also have

$$D_\phi(F_\Pi \| F_*) = \mathbb{E}[\phi(m_\Pi(U))] \le \frac{1}{|\Pi|}\sum_{\varpi\in\Pi}\mathbb{E}^{F_*}[\phi(m_0(\varpi U))] = \frac{1}{|\Pi|}\sum_{\varpi\in\Pi}\mathbb{E}^{F_*}[\phi(m_0(U))] \le \delta$$

because $F_0 \in \mathcal{N}_\delta$. This proves $F_\Pi \in \mathcal{N}_\delta$. Finally, as $\Pi$ is a group, for any function $h$ and any $\varpi \in \Pi$, we have

$$\sum_{\varpi'\in\Pi} h(\varpi' u) = \sum_{\varpi'\in\Pi} h(\varpi'\varpi^{-1}\varpi u) = \sum_{\varpi'\in\Pi} h(\varpi'\varpi u)\,.$$

Therefore, $m_\Pi(u) = m_\Pi(\varpi u)$ for each $\varpi \in \Pi$. It follows that for any $\varpi \in \Pi$ and $A \subset \mathcal{U}$,

$$\begin{aligned}\mathbb{E}^{F_\Pi}[\mathbb{1}\{\varpi U \in A\}] = \mathbb{E}^{F_*}[m_\Pi(U)\mathbb{1}\{\varpi U \in A\}] &= \mathbb{E}^{F_*}[m_\Pi(\varpi U)\mathbb{1}\{\varpi U \in A\}] \\ &= \mathbb{E}^{F_*}[m_\Pi(U)\mathbb{1}\{U \in A\}] = \mathbb{E}^{F_\Pi}[\mathbb{1}\{U \in A\}]\,,\end{aligned}$$

by $\Pi$-invariance of $F_*$. Hence, $F_\Pi$ is $\Pi$-invariant and so $F_\Pi \in \mathcal{N}_\delta^\Pi$.

It remains to show that $F_\Pi$ yields the optimal value $\kappa_0$ satisfies (1). Note by $\Pi$-invariance of $F_*$ that

$$\begin{aligned}\mathbb{E}^{F_\Pi}[k(U)] &= \frac{1}{|\Pi|}\sum_{\varpi\in\Pi}\mathbb{E}^{F_*}[m_0(\varpi U)k(\varpi\varpi^{-1}U)] \\ &= \frac{1}{|\Pi|}\sum_{\varpi\in\Pi}\mathbb{E}^{F_*}[m_0(U)k(\varpi^{-1}U)] \\ &= \frac{1}{|\Pi|}\sum_{\varpi\in\Pi}\mathbb{E}^{F_*}[m_0(U)k(\varpi U)] = \mathbb{E}^{F_0}[k^\Pi(U)] = \kappa_0\,,\end{aligned}$$

where the first equality on the final line is because $\{\varpi^{-1} : \varpi \in \Pi\} = \Pi$. An identical argument shows that $\mathbb{E}^{F_\Pi}[g_1(U)] = \mathbb{E}^{F_0}[g_1^\Pi(U)] \le P_1, \ldots, \mathbb{E}^{F_\Pi}[g_4(U)] = \mathbb{E}^{F_0}[g_4^\Pi(U)] = 0$. ∎

**Proof of Proposition A.2.** The minimization problem is additively separable across each $x \in \mathcal{X}$. The result follows by applying Proposition 2.1 for each $x$. ■

**Proof of Proposition A.3.** Follows by similar arguments to the proof of Proposition 2.1. ■

## G.7 Proofs for Appendix B

**Proof of Lemma B.1.** We prove the result only for $\underline{K}_\infty$; the result for $\overline{K}_\infty$ follows similarly. 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 $F_\varepsilon$ 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 Appendix G.2. Dropping dependence of $k$ and $g$ on $(\theta, \gamma)$, consider

$$\inf_F \mathbb{E}^F[k(U)] \quad \text{subject to} \quad \mathbb{E}^F[g_1(U)] \leq P_1\,, \ldots, \mathbb{E}^F[g_4(U)] = 0. \tag{A.21}$$

Identify each $F \in \mathcal{N}_\infty$ with its Radon–Nikodym derivative $m = \mathrm{d}F/\mathrm{d}F_* \in \mathcal{L}$ (see Appendix F). 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

$$\mathbb{I}_{C_2}(y_1, y_2) = \left[ \begin{array}{cl} 0 & \text{if } y_1 = 0, \text{ and } y_2 \in \mathbb{R}^{d_1}_- \times \{0_{d_2}\} \times \mathbb{R}^{d_3}_- \times \{0_{d_4}\}, \\ +\infty & \text{otherwise,} \end{array} \right.$$

and $\mathbb{I}_{C_+} : \mathcal{L} \to \mathbb{R} \cup \{+\infty\}$ is given by

$$\mathbb{I}_{C_+}(m) = \left[ \begin{array}{cl} 0 & \text{if } m \in \mathcal{L}_+, \\ +\infty & \text{otherwise.} \end{array} \right.$$

The *primal* problem for $y \in \mathbb{R}^{d+1}$ is $\min_{m \in \mathcal{L}} \varphi_\infty(m, y)$ and its *value* is $v_\infty(y) = \inf_{m \in \mathcal{L}} \varphi_\infty(m, y)$. By similar arguments to Lemmas G.1 and G.2, one may deduce that $\varphi_\infty$ is proper and convex and that $v_\infty : \mathbb{R}^{d+1} \to \mathbb{R} \cup \{+\infty\}$ is convex and its effective domain is $\mathcal{Y}_2$ (see display (A.7)).

Now consider the *dual* problem $\max_{y^\star \in \mathbb{R}^{d+1}} (y' y^\star - \varphi^\star_\infty(0, y^\star))$ at $y = 0$, where $\varphi^\star_\infty : \mathcal{E} \times \mathbb{R}^{d+1} \to \mathbb{R} \cup \{+\infty\}$ is the convex conjugate of $\varphi_\infty$. By direct calculation, with $y^\star = (y^\star_1, y^\star_2) \in \mathbb{R} \times \mathbb{R}^d$

we have

$$\begin{aligned}\varphi_\infty^\star(0,y^\star) &= \sup_{(m,y)\in\mathcal{L}_+\times\mathbb{R}^{d+1}} \left(y'y^\star - \langle m,k\rangle - \mathbb{I}_{C_2}\left(\langle m,1\rangle - 1 + y_1, \langle m,g\rangle - \vec{P} + y_2\right)\right)\\ &= \sup_{m\in\mathcal{L}_+} \left(-y_1^\star(\langle m,1\rangle - 1) - y_2^{\star\prime}\left(\langle m,g\rangle - \vec{P}\right) - \langle m,k\rangle\right) + \mathbb{I}_{C_2^o}(y^\star),\end{aligned}$$

where $C_2^o = \mathbb{R}\times\Lambda$, $\mathbb{I}_{C_2^o}(y^\star) = 0$ if $y^\star \in C_2^o$ and $+\infty$ otherwise, and it suffices to optimize over $\mathcal{L}_+$ because $\mathbb{I}_{C_+}(m) = +\infty$ for $m \in \mathcal{L}\setminus\mathcal{L}_+$. Write $y^\star \in C_2^o$ as $y^\star = (\zeta,\lambda)$, where $\zeta\in\mathbb{R}$ and $\lambda\in\Lambda$. By Rockafellar and Wets (1998, Definition 14.59 and Theorem 14.60), we have

$$\begin{aligned}\varphi_\infty^\star(0,(\zeta,\lambda)) &= \sup_{m\in\mathcal{L}_+} \mathbb{E}^{F_*}\left[m(U)(-k(U) - \zeta - \lambda' g(U))\right] + \zeta + \lambda'\vec{P}\\ &= \mathbb{E}^{F_*}\left[\sup_{x\geq 0} x(-k(U) - \zeta - \lambda' g(U))\right] + \zeta + \lambda'\vec{P}\\ &= \left[\begin{array}{ll} \zeta + \lambda'\vec{P} & \text{if } \zeta + F_*\text{-ess}\inf(k+\lambda' g) \geq 0,\\ +\infty & \text{otherwise,}\end{array}\right.\end{aligned}$$

provided $(\zeta,\lambda)\in\mathbb{R}\times\Lambda$. The dual value at $y = 0$ is therefore

$$\begin{aligned}&\sup_{\zeta\in\mathbb{R},\lambda\in\Lambda} -\zeta - \lambda'\vec{P} \quad \text{subject to } \zeta + F_*\text{-ess}\inf(k+\lambda' g) \geq 0\\ &= \sup_{\lambda\in\Lambda: F_*\text{-ess}\inf(k+\lambda' g) > -\infty} \left(F_*\text{-ess}\inf(k+\lambda' g) - \lambda'\vec{P}\right). \qquad \text{(A.22)}\end{aligned}$$

Lemma G.4(ii) implies that $0 \in \mathrm{ri}(\mathcal{Y}_2)$ under Condition S. It then follows by Bonnans and Shapiro (2000, Propositions 2.147 and 2.148(iii)) that the primal and dual values (A.21) and (A.22) are equal and the set of dual solutions is nonempty. ∎

Lemma B.3 is proved by modifying results of Csiszár and Matúš (2012) from a setting with moment equality restrictions to one with moment 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 $(\theta,\gamma)$ in what follows to simplify notation. Define $\mathcal{M} = \{m \in L^1(\mu) : \int mg\,\mathrm{d}\mu \text{ is finite}\}$ and $\mathcal{M}_+ = \{m\in\mathcal{M} : m \geq 0\ \mu\text{-a.e.}\}$. Note $\mathrm{d}F/\mathrm{d}\mu \in \mathcal{M}_+$ if and only if $F\in\mathcal{F}_\theta$. For $y = (y_1,y_2)\in\mathbb{R}\times\mathbb{R}^d$, let

$$\mathcal{M}[y] = \left\{m\in\mathcal{M}_+ : \int m\,\mathrm{d}\mu = 1 + y_1\,,\ \int mg\,\mathrm{d}\mu = \vec{P} + y_2\right\},$$

where the second integration is element-wise. Define $\varphi_a, \varphi_b : \mathbb{R}^{d+1} \to \mathbb{R}\cup\{+\infty\}$ by

$$\varphi_a(y) = \inf_{m\in\mathcal{M}[y]} \int mk\,\mathrm{d}\mu\,, \qquad \varphi_b(y) = \left[\begin{array}{ll} 0 & \text{if } y\in\{0\}\times\mathbb{R}_-^{d_1}\times\{0_{d_2}\}\times\mathbb{R}_-^{d_3}\times\{0_{d_4}\},\\ +\infty & \text{otherwise,}\end{array}\right.$$

with the understanding that $\varphi_a(y) = +\infty$ if the infimum runs over an empty set. Both $\varphi_a$ and $\varphi_b$ are proper and convex (note that $\mu\text{-ess}\inf |k| < \infty$ ensures that $\varphi_a(y) > -\infty$ for all $y$ and $|\varphi_a(y)| < \infty$ whenever $\mathcal{M}[y] \neq \emptyset$).

Let $\mathcal{V} = \{\int (1, g)\, m \,\mathrm{d}\mu : m \in \mathcal{M}_+\}$. Note that $\operatorname{dom} \varphi_a = \mathcal{V} - (1, \vec{P})$. By Condition $\mathrm{S}_{np}$ and similar arguments to the proof of Lemma G.4(ii), we have that $(1, \vec{P}) \in \operatorname{ri}(\mathcal{V} + (\{0\} \times \mathcal{C}))$. By Rockafellar (1970, Corollary 6.6.2), we also have $\operatorname{ri}(\mathcal{V} + (\{0\} \times \mathcal{C})) = \operatorname{ri}(\mathcal{V}) + \operatorname{ri}(\{0\} \times \mathcal{C})$ and therefore $0 \in \operatorname{ri}(\operatorname{dom} \varphi_a) + \operatorname{ri}(\{0\} \times \mathcal{C})$. As such, there exists $v \in \mathbb{R}^{d+1}$ such that $v \in \operatorname{ri}(\operatorname{dom} \varphi_a)$ and $-v \in \operatorname{ri}(\{0\} \times \mathcal{C}) \equiv -\operatorname{ri}(\operatorname{dom} \varphi_b)$, so $\operatorname{ri}(\operatorname{dom} \varphi_a) \cap \operatorname{ri}(\operatorname{dom} \varphi_b)$ is nonempty. Then by Fenchel's Duality Theorem (Rockafellar, 1970, Theorem 31.1),

$$\underline{K}_{np}(\theta; \gamma, P) = \inf_{y \in \mathbb{R}^{d+1}} (\varphi_a(y) + \varphi_b(y)) = \sup_{y^\star \in \mathbb{R}^{d+1}} (-\varphi_a^\star(y^\star) - \varphi_b^\star(-y^\star)), \tag{A.23}$$

where $\varphi_a^\star$ and $\varphi_b^\star$ are the convex conjugates of $\varphi_a$ and $\varphi_b$. Write $y^\star = (\zeta, \lambda)$. Then

$$\varphi_b^\star(-(\zeta, \lambda)) = \left[ \begin{array}{ll} 0 & \text{if } -\lambda \in \Lambda, \\ +\infty & \text{otherwise,} \end{array} \right. \tag{A.24}$$

and with $y = (y_1, y_2) \in \mathbb{R} \times \mathbb{R}_d$,

$$\begin{aligned} -\varphi_a^\star((\zeta, \lambda)) &= \inf_{y \in \mathbb{R}^{d+1}} \inf_{m \in \mathcal{M}[y]} \left( -\zeta y_1 - \lambda' y_2 + \int m k \,\mathrm{d}\mu \right) \\ &= \inf_{y \in \mathbb{R}^{d+1}} \inf_{m \in \mathcal{M}[y]} \left( \zeta + \lambda' \vec{P} + \int (k - \zeta - \lambda' g)\, m \,\mathrm{d}\mu \right) \\ &= \inf_{m \in \mathcal{M}_+} \left( \zeta + \lambda' \vec{P} + \int (k - \zeta - \lambda' g)\, m \,\mathrm{d}\mu \right). \end{aligned}$$

Let

$$Q(u, m(u)) = \left[ \begin{array}{ll} k(u) m(u) & \text{if } m(u) \geq 0, \\ +\infty & \text{otherwise,} \end{array} \right.$$

so that

$$-\varphi_a^\star((\zeta, \lambda)) = \inf_{m \in \mathcal{M}} \left( \zeta + \lambda' \vec{P} + \int Q(u, m(u)) - \big(\zeta + \lambda' g(u)\big)\, m(u) \,\mathrm{d}\mu(u) \right).$$

By Remark A.3 and Theorem A.4 of Csiszár and Matúš (2012), we have

$$\begin{aligned} -\varphi_a^\star((\zeta, \lambda)) &= \zeta + \lambda' \vec{P} + \int \inf_{x \geq 0} \big(k(u) - \zeta - \lambda' g(u)\big)\, x \,\mathrm{d}\mu(u) \\ &= \left[ \begin{array}{ll} -\infty & \text{if } \mu\text{-ess}\inf(k - \zeta - \lambda' g) < 0, \\ \zeta + \lambda' \vec{P} & \text{otherwise.} \end{array} \right. \end{aligned} \tag{A.25}$$

It now follows from (A.23), (A.24), and (A.25) that

$$\underline{K}_{np}(\theta;\gamma,P) = \sup_{\zeta\in\mathbb{R},\lambda\in\Lambda:\mu\text{-ess}\inf(k-\zeta+\lambda' g)\geq 0} \zeta - \lambda'\vec{P}$$
$$= \sup_{\lambda\in\Lambda:\mu\text{-ess}\inf(k+\lambda' g)>-\infty} \mu\text{-ess}\inf(k+\lambda' g) - \lambda'\vec{P}\,,$$

as required. ■

## G.8 Proofs for Appendix C

**Proof of Theorem C.1.** To simplify notation, we drop dependence of $g(u,\theta,\gamma_0,P_{20})$ on $(\gamma_0,P_{20})$ and $k(u,\theta,\gamma_0)$ on $\gamma_0$. Under the stated regularity conditions, $k$ and each entry of $g$ belong to $L^2(F_*)$ for all $\theta$ in a neighborhood of $\theta_*$.

Step 1: We first show $s \geq 2\mathbb{E}^{F_*}[\iota(U)^2]$. Let $L^2_0(F_*) = \{b \in L^2(F_*) : \mathbb{E}^{F_*}[b(U)] = 0\}$. Define $\mathbb{M} : L^2_0(F_*) \to L^2_0(F_*)$ by

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

Fix $b \in L^2_0(F_*)$. By Example 3.2.1 of Bickel, Klaassen, Ritov, and Wellner (1993), define a smooth parametric family $\{F_t : t \in (-1,1)\}$ passing through $F_*$ at $t=0$ via

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

Fix any $d_\theta \times (d_2 + d_4)$ matrix $A$ of full rank. 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 $\kappa(F_t) = \mathbb{E}^{F_t}[k(U,\theta(F_t))]$ and $\tilde{\iota}(u) = k_*(u) - \kappa_* - J'(AG)^{-1}Ag_*(u)$, we have

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

where the final line is because $\mathbb{M}$ is an orthogonal projection. For any $A$, we have

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

Hence,

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

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 $\mathbb{E}^{F_*}[(\mathbb{M}b(U))^2] \neq 0$, it follows from the preceding two displays that 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 we trivially have $s \geq 2\mathbb{E}^{F_*}[\iota(U)^2]$. Otherwise, choosing $b = \iota$ (which is valid because $\mathbb{E}^{F_*}[\iota(U)] = 0$ by construction) yields $s \geq 2\mathbb{E}^{F_*}[\iota(U)^2]$.

Step 2: We prove $s \leq 2\mathbb{E}^{F_*}[\iota(U)^2]$ by contradiction. Suppose there exists a sequence $\delta_n \downarrow 0$ 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

$$\frac{(\mathbb{E}^{\overline{F}_n}[k(U,\overline{\theta}_n)] - \mathbb{E}^{\underline{F}_n}[k(U,\underline{\theta}_n)])^2}{4\delta_n} \geq 2\mathbb{E}^{F_*}[\iota(U)^2] + \varepsilon\,. \tag{A.26}$$

As $\Theta$ is compact, (taking a subsequence if necessary) we have $\underline{\theta}_n \to \underline{\theta}^* \in \Theta$ and $\overline{\theta}_n \to \overline{\theta}^* \in \Theta$.

The spaces $\mathcal{L}$ and $\mathcal{E}$ with $\phi(x) = \frac{1}{2}(x-1)^2$ are equivalent to $L^2(F_*)$. 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 $\underline{m}_n = \mathrm{d}\underline{F}_n/\mathrm{d}F_*$ and $\overline{m}_n = \mathrm{d}\overline{F}_n/\mathrm{d}F_*$. As $\underline{F}_n, \overline{F}_n \in \mathcal{N}_{\delta_n}$,

$$\|\underline{m}_n - 1\|_2^2, \|\overline{m}_n - 1\|_2^2 \leq 2\delta_n \downarrow 0 \quad \text{as } n \to \infty. \tag{A.27}$$

By similar arguments to Lemma E.2, we may deduce $\mathbb{E}^{F_*}[g(U,\underline{\theta}^*)] = \mathbb{E}^{F_*}[g(U,\overline{\theta}^*)] = 0$. It then follows by identifiability of $\theta_*$ that $\underline{\theta}^* = \overline{\theta}^* = \theta_*$.

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

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

It follows by Cauchy–Schwarz and the fact that $G$ has full rank that $\|\underline{\theta}_n - \theta_*\| = O(\|\underline{m}_n - 1\|_2)$. Therefore, by (A.27), Cauchy–Schwarz, and $L^2(F_*)$ continuity of $\theta \mapsto g(\cdot, \theta, \gamma_0, P_{20})$ at $\theta_*$,

$$-G(\underline{\theta}_n - \theta_*) = \mathbb{E}^{F_*}[(\underline{m}_n(U) - 1)g_*(U)] + o(\delta_n^{1/2}) \tag{A.28}$$

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

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

However, by (A.28) and definition of $\mathbb{M}$ we also have

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

hence

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

with an analogous result holding for $\overline{m}_n$ and $\overline{\theta}_n$. It follows from the preceding display and its counterpart for $\overline{m}_n$ that

$$\frac{(\mathbb{E}^{\overline{F}_n}[k(U, \overline{\theta}_n)] - \mathbb{E}^{\underline{F}_n}[k(U, \underline{\theta}_n)])^2}{4\delta_n} = \frac{(\mathbb{E}^{F_*}[(\overline{m}_n(U) - \underline{m}_n(U))\iota(U)])^2}{4\delta_n} + o(1)\,. \tag{A.29}$$

Note that $\overline{m}_n \neq \underline{m}_n$ must hold for $n$ sufficiently large; otherwise, substituting (A.29) into (A.26) yields $o(1) \geq 2\mathbb{E}^{F_*}[\iota(U)^2] + \varepsilon$, a contradiction. By the triangle inequality and (A.27) we have

$$\|\overline{m}_n - \underline{m}_n\|_2^2 \leq 2\|\overline{m}_n - 1\|_2^2 + 2\|\underline{m}_n - 1\|_2^2 \leq 8\delta_n. \tag{A.30}$$

It follows by substituting (A.29) and (A.30) into (A.26) that

$$2\mathbb{E}^{F_*}[\iota(U)^2] + \varepsilon \leq \frac{2(\mathbb{E}^{F_*}[(\overline{m}_n(U) - \underline{m}_n(U))\iota(U)])^2}{\|\overline{m}_n - \underline{m}_n\|_2^2} + o(1) \leq 2\mathbb{E}^{F_*}[\iota(U)^2] + o(1),$$

where the second inequality is by Cauchy–Schwarz. As $n \to \infty$, $\varepsilon$ dominates the $o(1)$ term and we obtain a contradiction. ■

**Lemma G.10** *Suppose that the conditions of Theorem C.1 hold, $(\hat{\theta}, \hat{\gamma}, \hat{P}_2) \to_p (\theta_*, \gamma_0, P_{20})$, and $\mathbb{E}^{F_*}[g(U, \theta, \gamma, P_2)g(U, \theta, \gamma, P_2)']$, $\mathbb{E}^{F_*}[g(U, \theta, \gamma, P_2)k(U, \theta, \gamma)]$, $\mathbb{E}^{F_*}[g(U, \theta, \gamma, P_2)]$, $\mathbb{E}^{F_*}[k(U, \theta, \gamma)]$, $\frac{\partial}{\partial \theta'}\mathbb{E}^{F_*}[g(U, \theta, \gamma, P_2)]$, $\frac{\partial}{\partial \theta'}\mathbb{E}^{F_*}[k(U, \theta, \gamma)]$, and $\mathbb{E}^{F_*}[k(U, \theta, \gamma)^2]$ are all continuous in $(\theta, \gamma, P_2)$ at $(\theta_*, \gamma_0, P_{20})$. Then: $\hat{s} \to_p s$.*

**Proof of Lemma G.10.** Immediate by consistency of $(\hat{\theta}, \hat{\gamma}, \hat{P})$ and Slutsky's theorem. ■

# H Additional Details for Example 2.4

We first check Condition S'. Note $D_\phi(F\|F^*) = \frac{1}{2}\theta^2$ with $F$ the $N(\theta,1)$ distribution. Therefore, $|\theta| < \sqrt{2\delta}$ implies that Condition S' holds. In what follows we implicitly assume $|\theta| < \sqrt{2\delta}$.

By Proposition 2.1 and the discussion thereafter for KL neighborhoods,

$$\begin{aligned}
\overline{K}_\delta(\theta) &= \inf_{\eta>0,\lambda\in\mathbb{R}} \eta \log \mathbb{E}^{F_*}\left[e^{(\mathbb{1}\{U\le\theta\}-\lambda(U-\theta))/\eta}\right] + \eta\delta \\
&= \inf_{\eta>0,\lambda\in\mathbb{R}} \eta \log\left(e^{1/\eta}\int_{-\infty}^{\theta} e^{-\lambda(u-\theta)/\eta} f_N(u)\,du + \int_\theta^\infty e^{-\lambda(u-\theta)/\eta} f_N(u)\,du\right) + \eta\delta \\
&= \inf_{\eta>0,\lambda\in\mathbb{R}} \lambda\theta + \frac{\lambda^2}{2\eta} + \eta\log\left\{1 + (e^{1/\eta}-1)F_N\left(\theta+\frac{\lambda}{\eta}\right)\right\} + \eta\delta, \qquad \text{(A.31)}
\end{aligned}$$

where $f_N$ and $F_N$ denote the standard normal PDF and CDF, respectively, and the second line follows from the functional form of the moment generating function of the truncated normal distribution. As Condition S' holds at $\theta$, it follows from Proposition G.1 that minimizing values of $\eta$ and $\lambda$ exist. The first-order conditions (FOCs) are

$$\eta: \quad 0 = -\frac{\lambda^2}{2\eta^2} + \log\left\{1 + (e^{1/\eta}-1)F_N\left(\theta+\frac{\lambda}{\eta}\right)\right\} + \delta - \frac{1}{\eta}\frac{e^{1/\eta}F_N(\theta+\frac{\lambda}{\eta}) + \lambda(e^{1/\eta}-1)f_N(\theta+\frac{\lambda}{\eta})}{1+(e^{1/\eta}-1)F_N(\theta+\frac{\lambda}{\eta})}$$

$$\lambda: \quad 0 = \theta + \frac{\lambda}{\eta} + \frac{(e^{1/\eta}-1)f_N(\theta+\frac{\lambda}{\eta})}{1+(e^{1/\eta}-1)F_N(\theta+\frac{\lambda}{\eta})}.$$

The multiplier $\lambda$ enters both FOCs through the ratio $r := \lambda/\eta$. Rearranging the FOC for $\lambda$ yields

$$0 = (\theta + r) + (e^{1/\eta}-1)((\theta+r)F_N(\theta+r) + f_N(\theta+r)), \qquad \text{(A.32)}$$

which implicitly defines a function $r(\eta)$ on $(0,\infty)$. One may verify that $r(\eta)$ is strictly increasing, with $r(\eta) \to -\infty$ as $\eta \downarrow 0$. An asymptotic expansion of the error function (Gradshteyn and Ryzhik, 2014, formula 8.254) yields

$$F_N(x) = \frac{f_N(x)}{-x}\left(1 - \frac{1}{x^2} + O\left(\frac{1}{x^4}\right)\right), \quad \text{as } x \to -\infty, \qquad \text{(A.33)}$$

and hence

$$xF_N(x) + f_N(x) = \frac{f_N(x)}{x^2}\left(1 + O\left(\frac{1}{x^2}\right)\right), \quad \text{as } x \to -\infty. \qquad \text{(A.34)}$$

Substituting into (A.32) and taking logs, we obtain for small $\eta$ (and hence large negative $r$) that

$$\log(e^{1/\eta} - 1) = \frac{1}{2}(-\theta - r)^2 + \frac{1}{2}\log(2\pi) + 3\log(-\theta - r) + O\left(\frac{1}{(\theta + r)^2}\right).$$

As $\eta \log(e^{1/\eta} - 1) \to 1$ as $\eta \downarrow 0$ and $\log(-x)/x^2 \to 0$ as $x \to -\infty$, we obtain

$$-\theta - r = \sqrt{\frac{2}{\eta}}(1 + o(1)) \quad \text{as } \eta \downarrow 0. \tag{A.35}$$

Also note by (A.32), (A.33), and (A.34) that

$$(e^{1/\eta} - 1)F_N(\theta + r) = (\theta + r)^2 \left(1 + O\left(\frac{1}{(\theta + r)^2}\right)\right). \tag{A.36}$$

Substituting the FOC for $\lambda$ into the FOC for $\eta$ and rearranging, yields

$$\begin{aligned} 0 &= \frac{\theta\lambda}{\eta} + \frac{\lambda^2}{2\eta^2} + \log\left\{1 + (e^{1/\eta} - 1)F_N(\theta + \frac{\lambda}{\eta})\right\} + \delta - \frac{1}{\eta}\frac{e^{1/\eta}F_N\left(\theta + \frac{\lambda}{\eta}\right)}{1 + (e^{1/\eta} - 1)F_N\left(\theta + \frac{\lambda}{\eta}\right)} \\ &= -\frac{1}{2}\theta^2 + \log(e^{1/\eta} - 1) - \frac{1}{2}\log(2\pi) - \log(-\theta - r) + \delta - \frac{1}{\eta}\frac{F_N(\theta + r) + (e^{1/\eta} - 1)F_N(\theta + r)}{1 + (e^{1/\eta} - 1)F_N(\theta + r)}. \end{aligned}$$

Now substituting the approximation (A.36) into the previous display and using the fact that $\eta \log(e^{1/\eta} - 1) \to 1$ as $\eta \downarrow 0$, we obtain

$$\begin{aligned} 0 &= -\frac{1}{2}\theta^2 + \frac{1}{\eta} + o(1) - \frac{1}{2}\log(2\pi) - \log(-\theta - r) + \delta - \frac{1}{\eta}\left(1 - \frac{1}{(\theta + r)^2} + o\left(\frac{1}{(\theta + r)^2}\right)\right) \\ &= -\frac{1}{2}\theta^2 + o(1) - \frac{1}{2}\log(2\pi) - \log(-\theta - r) + \delta + \frac{1}{\eta(\theta + r)^2} + o\left(\frac{1}{\eta(\theta + r)^2}\right). \end{aligned}$$

Note by (A.35) that $\eta(\theta + r)^2 \to 2$ as $\eta \downarrow 0$. It follows that as $\delta \to \infty$, all terms in the above display remain bounded aside from $\log(-\theta - r)$. Therefore, the optimal $r = \lambda/\eta$ behaves like

$$\log(-\theta - r) = \frac{1}{2}(1 - \theta^2) - \frac{1}{2}\log(2\pi) + \delta + o(1). \tag{A.37}$$

Equivalently, by (A.35) we have that the optimal $\eta$ behaves like

$$\log \eta = \log(4\pi) - (1 - \theta^2) - 2\delta + o(1). \tag{A.38}$$

Note, in particular, that this implies that the optimal $\eta$ is always positive but converges to zero at an exponential rate in $\delta$. This approximation is verified numerically in Figure 7.

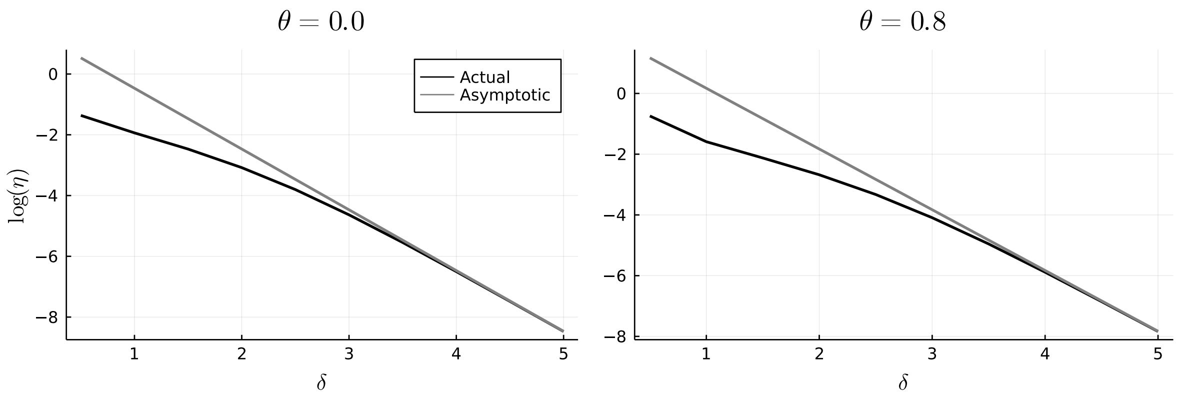


Figure 7: Log of the optimal value of $\eta$ ("actual") and its large-$\delta$ approximation $\log(4\pi) - (1-\theta^2) - 2\delta$ ("asymptotic").

Finally, substituting the FOC for $\lambda$ into (A.31) and then using (A.37), we obtain

$$\begin{aligned}\overline{K}_\delta(\theta) &= \eta\left(-\frac{1}{2}\theta^2 + \log(e^{1/\eta}-1) - \frac{1}{2}\log(2\pi) - \log(-\theta - r) + \delta\right) \\ &= \eta\left(-\frac{1}{2}\theta^2 + \log(e^{1/\eta}-1) - \frac{1}{2}\log(2\pi) - \frac{1}{2} + \frac{1}{2}\theta^2 + \frac{1}{2}\log(2\pi) - \delta + o(1) + \delta\right) \\ &= 1 - \frac{\eta}{2} + o(\eta)\,,\end{aligned}$$

It now follows by substituting (A.38) into the final line of the above display that

$$\overline{K}_\delta(\theta) = 1 - 2\pi e^{-2\delta-(1-\theta^2)}(1+o(1)).$$

This approximation is verified numerically in Figure 8.

It remains to derive an asymptotic expansion for the optimal value $\overline{\kappa}_\delta$. To this end, by the envelope theorem and the FOC for $\lambda$, we have

$$\frac{\partial \overline{K}_\delta(\theta)}{\partial\theta} = -\eta\theta,$$

where $\eta > 0$ solves the inner problem at $\theta$. It follows that $\overline{K}_\delta(\theta)$ is maximized at $\theta = 0$ for all $\delta > 0$, and therefore that $\overline{\kappa}_\delta = 1 - 2\pi e^{-2\delta-1}(1+o(1))$.

# 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.

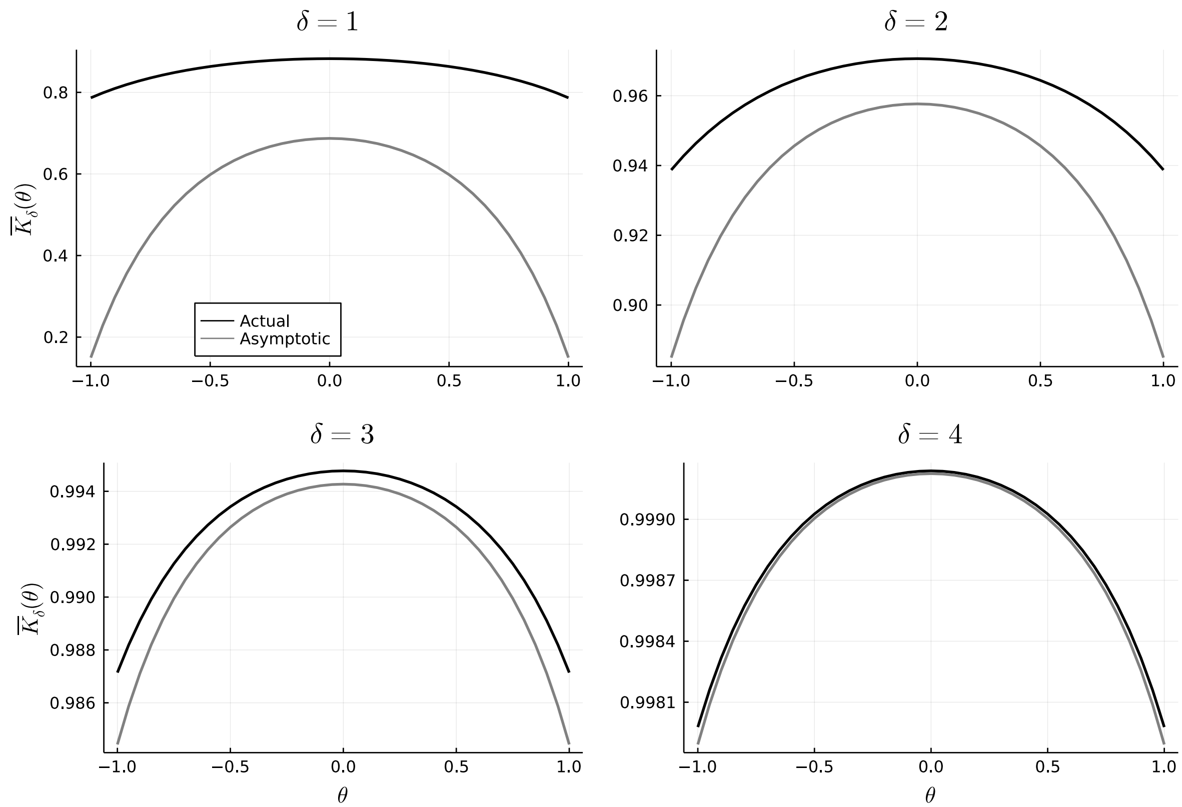


Figure 8: Objective function $\overline{K}_\delta(\theta)$ ("actual") and its large-$\delta$ approximation $1 - 2\pi e^{-2\delta-(1-\theta^2)}$ ("asymptotic").


Csiszár, I. and F. Matúš (2012). Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. *Kybernetika 48*(4), 637–689.

Gradshteyn, I. S. and I. M. Ryzhik (2014). *Table of Integrals, Series, and Products* (8 ed.). Academic Press.

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.