Inference in partially identified moment models via regularized optimal transport*

Grigory Franguridi[1] and Laura Liu[2]

[1]Center for Economic and Social Research, University of Southern California

[2]Department of Economics, University of Pittsburgh

December 24, 2025

Abstract

Partial identification often arises when the joint distribution of the data is known only up to its marginals. We consider the corresponding partially identified GMM model and develop a methodology for identification, estimation, and inference in this model. We characterize the sharp identified set for the parameter of interest via a support-function/optimal-transport (OT) representation. For estimation, we employ entropic regularization, which provides a smooth approximation to classical OT and can be computed efficiently by the Sinkhorn algorithm. We also propose a statistic for testing hypotheses and constructing confidence regions for the identified set. To derive the asymptotic distribution of this statistic, we establish a novel central limit theorem for the entropic OT value under general smooth costs. We then obtain valid critical values using the bootstrap for directionally differentiable functionals of Fang and Santos (2019). The resulting testing procedure controls size locally uniformly, including at parameter values on the boundary of the identified set. We illustrate its performance in a Monte Carlo simulation. Our methodology is applicable to a wide range of empirical settings, such as panels with attrition and refreshment samples, nonlinear treatment effects, nonparametric instrumental variables without large-support conditions, and Euler equations with repeated cross-sections.

JEL Classification: C14, C21, C23

Keywords: entropic optimal transport, partial identification, sharp identified set, moment condition, panel data, attrition

*We are grateful to Tim Armstrong, Yanqin Fan, Toru Kitagawa, Chen Qiu, seminar participants at the University of Copenhagen and New Economic School, and conference participants at the California Econometrics Conference 2025, Aarhus Workshop in Econometrics VII, and CEME Young Econometricians Conference 2025. Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health and in part by the Social Security Administration under Award Number U01AG077280. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1 Introduction

Many quantities of interest in economics are not point-identified under realistic modeling choices. Two broad reasons can account for this: incomplete data, where relevant variables are only partially observed, and incomplete models, where economic theory does not pin down a unique data-generating mechanism. In such settings, the available data and modeling assumptions generally imply a set of parameter values rather than a unique point. The partial identification literature, such as the pioneering work by Manski (1990), studies what can be learned about economically relevant parameters when the available information is insufficient for point identification; see also the handbook by Manski (2003) and the survey by Tamer (2010).

Partial identification also often arises when the joint distribution of the data is unknown, even though the marginal distributions are observed. We consider a generalized method of moments (GMM) model where the parameter $\theta_0 \in \Theta \subset \mathbb{R}^k$ satisfies moment conditions

$$\mathbb{E}_{\pi_0}[\phi(X, Y, \theta_0)] = 0, \tag{1}$$

but the joint distribution π_0 of (X, Y) is only known to lie in the set of couplings $\Pi(\mu, \nu)$ with observable marginals μ and ν. To characterize the sharp identified set, fix a candidate θ and consider the set

$$\nu_\Pi(\theta) = \{\mathbb{E}_\pi[\phi(X, Y, \theta)] : \pi \in \Pi(\mu, \nu)\}$$

of all moment values generated by couplings consistent with the observed marginals. Then θ is compatible with the data if and only if $0 \in \nu_\Pi(\theta)$, or equivalently,

$$0 = \operatorname{dist}(0, \nu_\Pi(\theta)) = \min_{\pi \in \Pi(\mu,\nu)} \|\mathbb{E}_\pi[\phi(X, Y, \theta)]\| = \max_{\|u\| \le 1} \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi\left[u'\phi(X, Y, \theta)\right]. \tag{2}$$

The last equality follows by norm duality and Sion's minimax theorem: see Section 2.1 for details. The inner minimization is an optimal transport (OT) problem with linear costs $u'\phi(x, y, \theta)$, producing the coupling that minimizes the moment violation in direction u. The outer maximization identifies the worst-case direction. Thus, θ belongs to the sharp identified set exactly when this max-min value equals zero.

As a toy example, consider a randomized controlled trial (RCT) with potential outcomes $Y(0) \sim \mu$ and $Y(1) \sim \nu$. The share of units that benefit from treatment $\theta = \mathbb{P}_\pi(Y(1) \ge Y(0))$ is not point-identified since only the marginals μ, ν of π are observed. The moment function $\phi(Y(0), Y(1), \theta) = \mathbf{1}\{Y(1) \ge Y(0)\} - \theta$ is one-dimensional, and θ belongs to the sharp identified set $\Theta_{I,0}$ if and only if there exists a coupling $\pi \in \Pi(\mu, \nu)$ such that $\mathbb{E}_\pi[\phi(Y(0), Y(1), \theta)] = 0$. Equivalently, $\Theta_{I,0} = [\underline{\theta}, \overline{\theta}]$, where $\underline{\theta} = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{P}_\pi(Y(1) \ge Y(0))$ and $\overline{\theta} = \max_{\pi \in \Pi(\mu,\nu)} \mathbb{P}_\pi(Y(1) \ge Y(0))$. This logic also extends to vector-valued or implicitly defined parameters.

In this paper, we develop a complete methodology for identification, estimation, and inference in the OT-based partially identified GMM model (1). First, we characterize the sharp identified set for the parameter of interest using OT, as discussed above. The characterization provides both a geometric interpretation via support functions and informs an estimation procedure for the identified set.

Second, the implied estimation procedure involves solving an empirical version of the classical OT problem, which is known to be sensitive to sampling noise and having a slow convergence rate and a nonstandard limit distribution. To overcome this challenge, we employ *entropic regularization*, which penalizes the negative entropy of the joint distribution. Entropic OT has been widely used in statistics and machine learning since the seminal works by Cuturi (2013) and Galichon and Salanié (2022), see the literature review below. This regularization yields a strictly convex problem that restores the usual $\sqrt{n}$-convergence and asymptotic normality, albeit at a cost of introducing a regularization bias. It is also computationally attractive, admitting fast implementation via the *Sinkhorn algorithm* for the inner (entropic OT) problem and the projected gradient ascent for the outer problem in the max-min representation (2).

Third, we develop a procedure to test hypotheses and construct confidence regions for the identified set. To this end, we establish a uniform central limit theorem (CLT) for the entropic OT value uniformly in direction $u \in \mathbb{B}$ and parameter $\theta \in \Theta$. We build on Mena and Niles-Weed (2019) and Goldfeld et al. (2024) and extend their framework to accommodate arbitrary smooth cost functions. To the best of our knowledge, we are the first to establish a CLT of such generality for the entropic OT value.

We then apply the functional delta method to the max functional to obtain the asymptotic distribution of our test statistic, relying on a result by Franguridi and Moon (2025). This distribution depends on the unknown argmax set and is not available in closed form. Moreover, the standard bootstrap may be invalid when the hypothesized parameter value is on the boundary of the identified set, which occurs when the argmax set is not a singleton. We resolve this issue by employing the bootstrap for directionally differentiable functionals of Fang and Santos (2019). The resulting bootstrap-based test controls size locally uniformly and can be inverted to obtain a confidence region.

Finally, our estimator applies broadly to settings where one observes marginal distributions but lacks the knowledge of the joint distribution. In Section 4, we discuss four examples: fixed effects panel logit with attrition and refreshment samples, nonlinear treatment effects, nonparametric instrumental variables (IV) without a large support condition, and the Euler equation with repeated cross-sections. In the first example, we exploit the panel structure by fixing the joint distribution of retainers and solving OT only for attriters, which tightens the bounds for the common slope coefficient and average marginal effects (AME).

Related literature. First, our paper contributes to the extensive literature on partially identified models. Among others, Imbens and Manski (2004) and Stoye (2009) study confidence sets for partially identified parameters with uniform coverage and optimality properties, Beresteanu and Molinari (2008) and Bontemps et al. (2012) develop asymptotic theory and geometric characterizations for partially identified models using random sets and support functions, and Romano and Shaikh (2010) provide general procedures for inference on identified sets defined via minimization of a criterion function. For an overview of the broader literature on partial identification and inference, we refer to the review by Canay and Shaikh (2017). The closest antecedent is Beresteanu et al. (2011), which characterizes sharp identification regions in models with convex moment predictions using the theory of random sets and support-function representations. Our characterization of the identified set is in this spirit but does not rely on representations via random sets.

Second, there has been a growing literature that applied OT methods to economic problems. The paper closest to ours is Fan et al. (2025), which studies partial identification of a finite-dimensional parameter defined by a moment equality model with incomplete data via a classical OT representation of the identified set. In contrast, our paper employs entropic OT and also develops the methodology for estimation and inference, and hence goes beyond identification. More broadly, OT has been deployed in a variety of economic applications, such as discrete choice models (Chiong et al., 2016), covariate matching for causal effects (Gunsilius and Xu, 2021), nonlinear difference-in-differences for multivariate counterfactuals (Torous et al., 2024), policy learning in matching markets (Hazard and Kitagawa, 2025), and combining stated and revealed preferences (Meango et al., 2025). Other works include Voronin (2025), which introduces a generalized version of OT for estimation in a large class of partially identified models, and Schennach and Starck (2025), which uses OT for estimation and inference in the overidentified GMM with measurement errors.

Third, we employ entropic regularization for the classical OT, which yields a strictly convex program that can be solved efficiently by the Sinkhorn algorithm. Entropic OT is now widely used in high-dimensional statistics and machine learning. In seminal works, Cuturi (2013) introduces Sinkhorn distances as a fast approximation to Wasserstein distances based on entropic regularization, and Galichon and Salanié (2022) show how entropic regularization of social surplus yields efficient algorithms for estimating matching models. We also contribute to the statistical theory of the entropic OT and rely on two important prior works. Mena and Niles-Weed (2019) establish asymptotic normality of the entropic OT cost for quadratic cost functions, and Goldfeld et al. (2024) extend this analysis using empirical process theory to derive semiparametric efficiency bounds and bootstrap inference procedures. Our CLT extends these results by allowing for arbitrary smooth cost functions, which is essential when the cost itself encodes economically meaningful restrictions.

Finally, our analysis for the panel logit with attrition and refreshment is closely related to recent contributions on panel surveys with nonignorable attrition and refreshment in Franguridi et al.

(2025a,b). They develop computationally feasible and robust procedures for estimation and inference in this setting under the assumption of additively separable attrition that restores point identification. In contrast, we do not impose any assumptions on attrition, and hence our model is partially identified even with refreshment samples. We then show how to characterize and estimate bounds on the common slope parameter and the AME (see, e.g., Davezies et al. (2024)), as well as conduct inference in this setting.

The remainder of the paper is organized as follows. Section 2 introduces the partially identified OT-based GMM model, and characterizes the sharp identified set for the parameter of interest. Section 3 develops procedures for estimation and inference using entropic regularization. Section 4 discusses several economic models that fit our general framework. Section 5 conducts a Monte Carlo simulation for the fixed effects panel logit with attrition and refreshment. Section 6 concludes. Appendix A contains proofs of all the theoretical results, and Appendix B provides additional details for the fixed effects panel logit with attrition and refreshment.

2 Setup and partial identification

2.1 Model and sharp identified set

Let X and Y be random vectors in $\mathbb{R}^d$ with the joint distribution π_0 identified only up to the class $\Pi(\mu, \nu)$ of joint distributions with marginals μ and ν. The true value θ_0 of a parameter of interest $\theta \in \Theta \subset \mathbb{R}^k$ uniquely satisfies the set of moment conditions

$$\mathbb{E}_{\pi_0}[\phi(X, Y, \theta_0)] = 0,$$

where ϕ is a p-dimensional moment function. Since π_0 is only partially identified, so is the parameter θ in general. We allow for arbitrary $\dim(\phi) = p$ and $\dim(\theta) = k$ as long as the identified set is nonempty and compact. For $p > k$, the additional moments introduce extra directions to detect violations and may tighten the identified set.

For each fixed parameter value θ, consider the set of *moment predictions*

$$\nu_\Pi(\theta) = \{\mathbb{E}_\pi \, \phi(X, Y, \theta) : \pi \in \Pi\},$$

i.e., the set of all moment vectors consistent with the point-identified set of distributions $\Pi = \Pi(\mu, \nu)$. Because expectation is linear and Π is convex, $\nu_\Pi(\theta)$ is convex. The sharp identified set is then

$$\Theta_{I,0} = \{\theta \in \Theta : 0 \in \nu_\Pi(\theta)\} = \{\theta \in \Theta : D_0(\theta) = 0\},$$

where $D_0(\theta) = d(0, \nu_\Pi(\theta))$ denotes the Euclidean distance from the origin to the set $\nu_\Pi(\theta)$. Identification is equivalent to the origin lying in the set of moment predictions (Beresteanu et al., 2011).

Note that, although $\nu_\Pi(\theta)$ is convex for each θ, the identified set $\Theta_{I,0}$ need not itself be convex.

Let $\mathbb{B}$ be the unit ball in $\mathbb{R}^p$. By norm duality and Sion's minimax theorem, the distance $D_0(\theta)$ admits the following max-min representation

$$\begin{aligned} D_0(\theta) = d\left(0, \nu_\Pi(\theta)\right) &= \min_{\pi\in\Pi} \left\|\mathbb{E}_\pi\, \phi(X,Y,\theta)\right\|_2 = \min_{\pi\in\Pi} \max_{u\in\mathbb{B}} \mathbb{E}_\pi \left[u'\phi(X,Y,\theta)\right] \\ &= \max_{u\in\mathbb{B}} \underbrace{\min_{\pi\in\Pi} \mathbb{E}_\pi \left[u'\phi(X,Y,\theta)\right]}_{\text{OT with cost } u'\phi} =: \max_{u\in\mathbb{B}} c_\theta(u). \end{aligned}$$

The inner minimization over couplings π is an OT problem whose cost is the directional moment $u'\phi$. The outer maximization selects the direction u in which the model slackness is largest. Since $D_0(\theta) = \max_{u\in\mathbb{B}} c_\theta(u)$ and $\theta \in \Theta_{I,0}$ if and only if $D_0(\theta) = 0$, we have that θ belongs to the sharp identified set exactly when $\max_{u\in\mathbb{B}} c_\theta(u) = 0$, or equivalently, when $c_\theta(u) \le 0$ for all $u \in \mathbb{B}$. This max-min representation informs our estimation and inference procedure based on estimating $c_\theta(u)$ for $u \in \mathbb{B}$ and checking whether the maximum is close to zero: see Section 3.

Remark 1 (Geometric interpretation of $c_\theta(u)$ and $D_0(\theta)$)**.** *Since $\nu_\Pi(\theta)$ is convex, $\theta \in \Theta_{I,0}$ (i.e., $0 \in \nu_\Pi(\theta)$) holds if and only if no direction $u \in \mathbb{B}$ separates the origin from $\nu_\Pi(\theta)$, which is equivalent to*

$$\min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi \left[u'\phi(X,Y,\theta)\right] \le 0, \quad \forall\, u \in \mathbb{B}.$$

In terms of the OT value function $c_\theta(u)$, this is exactly the condition $c_\theta(u) \le 0$ for all $u \in \mathbb{B}$. For a different use of separating hyperplane ideas to characterize identified sets, see Botosaru et al. (2024). Moreover, by convex duality, the negative distance admits the support function representation

$$-D_0(\theta) = \min_{u\in\mathbb{B}} \underbrace{\max_{\pi\in\Pi(\mu,\nu)} u'\, \mathbb{E}_\pi\, \phi(X,Y,\theta)}_{\textit{support function of } \nu_\Pi(\theta)} = \min_{u\in\mathbb{B}} \max_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi \left[u'\phi(X,Y,\theta)\right].$$

These identities clarify the connection to models with convex moment predictions in Beresteanu et al. (2011), while highlighting the key distinction that our direction-indexed inner problem defining $c_\theta(u)$ is infinite-dimensional.

Note that our setup is not a standard moment inequality model because the inner minimization over π depends on the direction u and delivers a direction-dependent evaluation of the set of predicted moments rather than a fixed collection of inequalities. It is also different from intersection bounds that aggregate separate scalar constraints.

To make the informal derivation above rigorous, we impose the following mild assumptions to ensure that the sharp identified set is nonempty and compact. Both of these properties are crucial for the Hausdorff consistency of the associated estimator, see Theorem 2.

Assumption 1. *(i) The parameter space $\Theta \subset \mathbb{R}^k$ is nonempty and compact.*

(ii) The distributions μ and ν have compact supports $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^d$, respectively.

(iii) The identified set $\Theta_{I,0}$ is nonempty, i.e., there exists $\theta_0 \in \Theta$ and $\pi_0 \in \Pi(\mu, \nu)$ such that $\mathbb{E}_{\pi_0}[\phi(X, Y, \theta_0)] = 0$.

(iv) For each $\theta \in \Theta$, the function $(x, y) \mapsto \phi(x, y, \theta)$ is continuous.

(v) For each $\pi \in \Pi(\mu, \nu)$, the function $\theta \mapsto \mathbb{E}_\pi[\phi(X, Y, \theta)]$ is continuous.

Theorem 1 (characterization of identified set)**.** *Suppose Assumption 1 holds. Then the identified set $\Theta_{I,0}$ is nonempty and compact, and $\Theta_{I,0} = \{\theta \in \Theta : D_0(\theta) = 0\}$.*

Proof. See Appendix A.1. □

2.2 Toy example

To illustrate the characterization above, let us revisit the toy example in the introduction. Consider a RCT with potential outcomes $Y(0) \sim N(0, 1)$ and $Y(1) \sim N(2, 1)$. The parameter of interest is the share of units that benefit from treatment,

$$\theta = \mathbb{P}_\pi\left(Y(1) > Y(0)\right) = \mathbb{E}_\pi\left[\mathbf{1}\{Y(1) > Y(0)\}\right].$$

The corresponding scalar moment condition is $\mathbb{E}_\pi\left[\phi\left(Y(0), Y(1), \theta\right)\right] = 0$, where $\phi\left(Y(0), Y(1), \theta\right) = \mathbf{1}\{Y(1) > Y(0)\} - \theta$. The identified set is then

$$\Theta_{I,0} := \left\{\theta \in \mathbb{R} : \max_{u \in [-1,1]} c_\theta(u) = 0\right\}, \quad \text{where } c_\theta(u) = \min_{\pi \in \Pi} \mathbb{E}_\pi\left[u\phi(Y(0), Y(1), \theta)\right].$$

This characterization has a simple interpretation as a two-sided game. Since ϕ is scalar, u simply flips the sign of the moment: $u = 1$ tests whether $\mathbb{E}_\pi[\phi\left(Y(0), Y(1), \theta\right)]$ is positive under some $\pi \in \Pi$, while $u = -1$ tests the opposite. The adversary chooses the least favorable coupling under each sign. Thus,

$$\max_{u \in \{-1,0,1\}} c_\theta(u) = \max\left\{0,\ \min_{\pi \in \Pi} \mathbb{P}_\pi(Y(1) > Y(0)) - \theta,\ \ \theta - \min_{\pi \in \Pi} \mathbb{P}_\pi(Y(1) > Y(0))\right\}$$

is the worst of these two one-sided checks, and equals 0 at $u = 0$. If neither side can produce a positive value, then θ is in the identified set $\Theta_{I,0}$.

For Gaussian marginals with common variance σ^2 and means $\mu_1 \geq \mu_0$, the classical sharp bounds

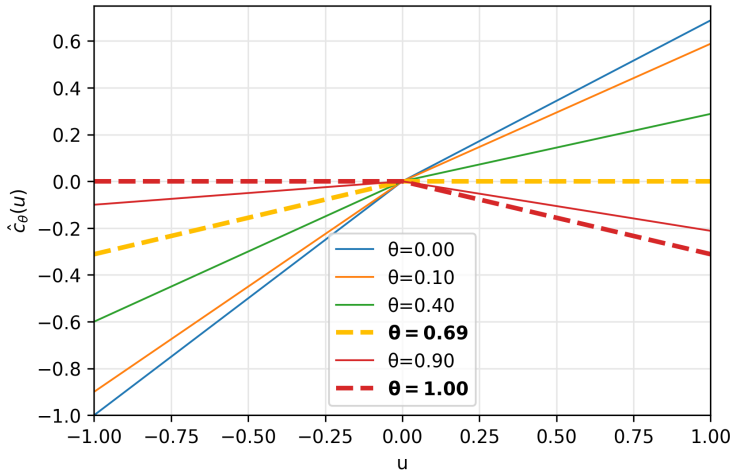

Figure 1: Population cost function $u \mapsto c_\theta(u)$ for the RCT toy example

of Makarov (1982) yield

$$1 - 2\Phi\left(\frac{-(\mu_1 - \mu_0)}{2\sigma}\right) \leq \mathbb{P}_\pi\left(Y(1) \geq Y(0)\right) \leq \ 1,$$

as also discussed in Firpo and Ridder (2019). With $\sigma = 1$, $\mu_0 = 0$, and $\mu_1 = 2$, the identified set is $0.69 \leq \mathbb{P}_\pi\left(Y(1) \geq Y(0)\right) \leq 1$, which matches our characterization.

Figure 1 plots $u \mapsto c_\theta(u)$ for $u \in [-1, 1]$. The curve is piecewise linear, anchored at $u = 1$ by the lowest feasible beneficiary share minus θ ($\min_{\pi \in \Pi} \mathbb{P}_\pi(Y(1) > Y(0)) - \theta = 0.69 - \theta$), and at $u = -1$ by θ minus the highest feasible beneficiary share ($\theta - \max_{\pi \in \Pi} \mathbb{P}_\pi(Y(1) > Y(0)) = \theta - 1$). The identification check reduces to verifying whether this curve stays weakly below zero. For $\theta < 0.69$, the right endpoint is above zero, whereas for $\theta \in [0.69, 1]$, the whole curve remains nonpositive, and hence $\Theta_{I,0} = [0.69, 1]$.

3 Estimation and inference

Suppose now that we have access to random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from distributions μ and ν, respectively. For simplicity, we let $n = m$ and assume that the two samples are independent, although these assumptions can be relaxed at the expense of heavier notation. The goal of this section is to describe our estimator of the sharp identified set Θ_I and a testing procedure for the hypothesis $H_0 : \theta = \theta_0$.

3.1 Estimation

The characterization of the sharp identified set in Theorem 1 directly informs an estimation procedure. Denote by $\hat{\mu}$, $\hat{\nu}$ the empirical distributions based on samples (X_i) and (Y_j), and let $\hat{\Pi} = \Pi(\hat{\mu}, \hat{\nu})$ be the set of joint distributions with marginals $\hat{\mu}$ and $\hat{\nu}$. The sample analog of $c_{\theta,0}(u)$ is then

$$\hat{c}_{\theta,0}(u) = \min_{\pi \in \hat{\Pi}} \mathbb{E}_\pi \left[u' \phi(X, Y, \theta) \right].$$

This is the value of the empirical OT problem with the cost function $(x, y) \mapsto u'\phi(x, y, \theta)$. This is an infinite-dimensional linear program that is known to be computationally challenging, sensitive to sampling noise, and having nonstandard convergence rates when the (effective) dimensions of X and Y are greater than 4, see, e.g., Cuturi (2013); Hundrieser et al. (2024). We therefore suggest using *entropic regularization* – a classical technique for improving analytical and computational properties of OT that was introduced in Cuturi (2013) and a working paper version of Galichon and Salanié (2022). This constitutes adding a term to the cost function that penalizes deviations from the independence distribution $\hat{\mu} \otimes \hat{\nu}$, viz.,

$$\hat{c}_{\theta,\varepsilon}(u) = \min_{\pi \in \hat{\Pi}} \mathbb{E}_\pi \left[u' \phi(X, Y, \theta) \right] + \varepsilon \cdot \mathrm{KL}(\pi \,||\, \hat{\mu} \otimes \hat{\nu}),$$

where KL is the Kullback-Leibler divergence,

$$\mathrm{KL}(\pi \,||\, \hat{\mu} \otimes \hat{\nu}) = \int \log \frac{d\pi}{d(\hat{\mu} \otimes \hat{\nu})}(x, y)\, d\pi(x, y).$$

The advantage of such regularization is that the program becomes strictly convex, much less sensitive to sampling noise, and regains standard asymptotics as we show in the next subsection. Importantly, the regularized program can be solved using explicit iterations of the Sinkhorn algorithm, see Section 3.3. Although regularization introduces (small) bias, explicit debiasing procedures are available in the literature, see, e.g., Pooladian et al. (2022) and references therein.

The identified set of interest then becomes

$$\Theta_{I,\varepsilon} = \{\theta \in \Theta : \, D_\varepsilon(\theta) = 0\},$$

where $\varepsilon > 0$ is a penalty parameter and

$$D_\varepsilon(\theta) = \max_{u \in \mathbb{B}} c_{\theta,\varepsilon}(u)$$

is the population distance statistic. Throughout the rest of the paper, we drop the subscript ε for brevity.

We define our estimator of Θ_I as

$$\widehat{\Theta}_I = \left\{\theta \in \Theta : \hat{D}(\theta) \leq \eta_n\right\},$$

where $\eta_n > 0$ is a tuning parameter and the (sample) distance statistic is

$$\hat{D}(\theta) = \max_{u \in \mathbb{B}} \hat{c}_\theta(u). \tag{3}$$

We impose the following assumptions.

Assumption 2. *(i) There exists a nondecreasing function $m : \mathbb{R}_+ \to \mathbb{R}_+$ such that $m(0) = 0$, $m(\delta) > 0$ for $\delta > 0$ and $D(\theta) \geq m(d(\theta, \Theta_I))$ for all $\theta \in \Theta$.*

(ii) $P(\|\hat{D} - D\|_\infty \leq r_n) \to 1$ for a deterministic sequence $r_n \downarrow 0$.

(iii) $\eta_n \downarrow 0$ and $r_n = o(\eta_n)$.

Assumption 2(i) imposes weak separation of the identified set by the criterion function. Theorem 3 below implies that Assumption 2(ii) holds with $r_n = Cn^{-1/2}$ for some constant C. Finally, Assumption 2(iii) requires picking η_n that converges to zero sufficiently slowly.

The following theorem establishes the convergence of our estimator to the sharp identified set in the Hausdorff distance d_H.

Theorem 2 (consistency)**.** *Under Assumptions 1 and 2, we have*

$$d_H(\widehat{\Theta}_I, \Theta_I) = o_p(1).$$

Proof. See Appendix A.2. □

3.2 Inference

Now our goal is to develop a test of the hypothesis $H_0 : \theta = \theta_0$. To this end, we use the rescaled distance $\sqrt{n} \cdot \hat{D}(\theta_0)$ as the test statistic. We characterize its asymptotic behavior by first establishing a novel CLT for the regularized OT value $\hat{c}_{\theta_0}(u)$, uniformly in $u \in \mathbb{B}$ and $\theta_0 \in \Theta$, and then applying the functional delta method to obtain the limiting distribution of our test statistic. Since this distribution does not have a simple form, we employ a result in Franguridi and Moon (2025) to establish the validity of the bootstrap for directionally differentiable functionals of Fang and Santos (2019).

Remark 2 (KS vs. CvM statistics)**.** *Our (population) statistic*

$$D(\theta) = \max_{u \in \mathbb{B}} c_\theta(u)$$

uses maximization that is characteristic of Kolmogorov-Smirnov (KS) type statistics in the moment inequality literature, see, e.g., Andrews and Shi (2013). The KS-type statistics are powerful against local alternatives with few violations (Armstrong, 2015, 2018), and provide diagnostics for the most binding inequality.

Alternatively, we could consider the Cramér-von Mises (CvM) type statistic

$$\tilde{D}(\theta) = \int_{\mathbb{B}} c_\theta(u)_+^2 \, d\omega(u),$$

for a probability measure ω on $\mathbb{B}$ with full support. The CvM-type statistics are powerful against diffused local alternatives (Andrews and Shi, 2013), retain power under weak identification (Bugni, 2010), typically exhibit smaller finite-sample size distortions, and are less sensitive to slack inequalities. While our methodology can be extended to the CvM type statistic, we focus on the KS type statistic in this paper since it is simple to implement and demonstrates good size and power performance in the Monte Carlo simulations.

Finally, yet another choice of a test statistic is a self-normalized moment violation statistic as in Chetverikov (2018). We leave consideration of such a statistic for future work.

Our first result is the uniform CLT for the regularized OT value, which we derive under the following assumptions.

Assumption 3. *(i) The probability measures μ and ν have bounded, convex supports $\mathcal{X}$ and $\mathcal{Y}$ in $\mathbb{R}^d$.*

(ii) For all $j = 1, \ldots, p$, the moment function $\phi_j : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is such that $\phi_j \in C^s(\mathcal{X} \times \mathcal{Y})$ with $s > d/2$.

(iii) The estimators $\hat{\mu}, \hat{\nu}$ are empirical measures based on independent random samples from μ and ν, respectively.

Assumption 3(i) helps establish uniform bounds on the optimal potentials and their derivatives. At the expense of more complicated proofs, this assumption can be relaxed to bounds on the tails of μ, ν related to the cost function, similar to Goldfeld et al. (2024). Assumption 3(iii) can be relaxed to accommodate dependence within samples, e.g., when the samples are stationary β-mixing processes, as well as dependence across samples, see Remark 11 in Goldfeld et al. (2024).

Theorem 3 (uniform CLT for regularized OT value)**.** *Suppose that Assumption 1(i) and Assumption 3 holds. Then there exists a tight Gaussian process $\mathbb{G}$ on $C(\mathbb{B} \times \Theta)$ such that*

$$\sqrt{n}\left(\hat{c}_\theta(u) - c_\theta(u)\right) \rightsquigarrow \mathbb{G}(u, \theta) \text{ in } C(\mathbb{B} \times \Theta).$$

Proof. See Appendix A.3. □

Remark 3. *If* $\dim \phi = 1$*, then considering this statement at a single point* $u = 1$ *and given* θ *delivers asymptotic normality of the regularized OT value under an arbitrary smooth cost function* ϕ*. To our knowledge, this is the first such result available in the literature. In our derivation, however, we relied heavily on the arguments in Mena and Niles-Weed (2019), who were the first to establish the asymptotic normality when the cost function is quadratic, and Goldfeld et al. (2024), who extended this result using empirical processes theory.*

Applying the functional delta method to this uniform convergence statement with the functional $\chi(c) := \max_{u \in \mathbb{B}} c(u)$ leads to the following result.

Corollary 1 (asymptotic distribution of the distance statistic)**.** *Suppose that Assumption 3 holds. Then, under the null hypothesis* $H_0 : \theta = \theta_0$*, we have*

$$\sqrt{n}(\hat{D}(\theta_0) - D(\theta_0)) = \sqrt{n}(\chi(\hat{c}_{\theta_0})) - \chi(c_{\theta_0}) = \max_{u \in U_c(\theta_0)} \mathbb{G}(u, \theta_0),$$

where $U_c(\theta_0) = \arg\max_{u \in \mathbb{B}} c_{\theta_0}(u)$*.*

Proof. See Appendix A.4. □

The asymptotic distribution of the distance statistic is neither available in closed form, nor is easy to simulate from, because it depends on the unknown features of the data-generating process such as the argmax set $U_c(\theta_0)$. Moreover, standard bootstrap often fails to control size uniformly in partially identified models (Andrews and Guggenberger, 2009; Andrews and Han, 2009). This failure occurs when the parameter is on the boundary of the parameter space (Andrews, 2000). In our model, this corresponds to the case where the hypothesized value θ_0 is on the boundary of the identified set or, equivalently, where $U_c(\theta_0)$ is not a singleton, such as $\theta \in \{0.69, 1.00\}$ in Figure 1.

To overcome this challenge, we make use of the bootstrap for directionally differentiable functionals of Fang and Santos (2019). Our testing procedure is described in Algorithm 1 and depends on an additional tuning parameter ι_n. We impose the following assumption.

Assumption 4 (bootstrap validity)**.** *(i) There exists* $\kappa > 0$ *such that*

$$c_{\theta_0}(u) \leq \max_{v \in \mathbb{B}} c_{\theta_0}(v) - \kappa \cdot d_H(u, U_c(\theta_0)) \text{ for all } u \in \mathbb{B}.$$

(ii) $\iota_n \downarrow 0$ *and* $n^{-1/2}\iota_n \uparrow \infty$*.*

Assumption 4(i) posits that the maxima of c_{θ_0} are well-separated. This assumption is equivalent to the (super)gradient of c_{θ_0} being bounded away from zero on the complement of the argmax set $U_c(\theta_0)$. It also suffices for this assumption to hold in a small neighborhood around $U_c(\theta_0)$ rather than on the entire ball $\mathbb{B}$. Assumption 4(ii) requires ι_n to converge to zero slower than $n^{-1/2}$. This

Algorithm 1 Testing $H_0 : \theta = \theta_0$

Require: θ_0, ε, ι_n, number of bootstrap draws B

Ensure: Test decision for H_0

Compute the support function estimate:

Compute $\hat{c}(u) = \hat{c}_{\theta_0}(u)$ using Sinkhorn algorithm with regularization parameter ε

Compute $\hat{D}(\theta_0) \leftarrow \max_{u \in \mathbb{B}} \hat{c}(u)$ via projected gradient ascent

Define the set

$$\hat{U}_n \leftarrow \left\{ u \in \mathbb{B} : \hat{c}(u) \geq \max_{\|v\| \leq 1} \hat{c}(v) - \iota_n \right\}$$

Bootstrap procedure:

for $b = 1, \ldots, B$ **do**

  Draw bootstrap samples $X_1^b, \ldots, X_n^b$ and $Y_1^b, \ldots, Y_n^b$ with replacement from $\hat{\mu}$ and $\hat{\nu}$

  Compute $\hat{c}^{*,b}(u)$, the regularized OT value on $\{X_i^b, Y_i^b\}_{i=1}^n$, using Sinkhorn algorithm

  Compute the bootstrap statistic

$$\hat{T}^{*,b} \leftarrow \max_{u \in \hat{U}_n} \sqrt{n} \left(\hat{c}^{*,b}(u) - \hat{c}(u) \right)$$

end for

Compute critical value and test decision:

Let $\hat{q}_\alpha^*$ be the $(1-\alpha)$ empirical quantile of $\{\hat{T}^{*,1}, \ldots, \hat{T}^{*,B}\}$

return Reject H_0 if $\sqrt{n}\, \hat{D}(\theta_0) > \hat{q}_\alpha^*$

guarantees that the enlarged argmax $\hat{U}_n$ converges in the Hausdorff distance to the true argmax $U_c(\theta_0)$. If $U_c(\theta_0)$ is known to be a singleton, we can take $\hat{U}_n$ to be the sample argmax of $\hat{c}_{\theta_0}$.

It is straightforward to establish that under Assumption 4, our testing procedure controls size locally uniformly in the sense of Corollary 3.2 of Fang and Santos (2019). We refer the reader to Theorem 4 of Franguridi and Moon (2025) for details.

Remark 4. *It is possible to make our test control size over the original identified set $\Theta_{I,0}$ by reducing the value of the distance statistic appropriately. Namely, as implied by the proof of Proposition 2 in Hazard and Kitagawa (2025),*

$$0 \leq \hat{c}_{\theta_0,\varepsilon}(u) - \hat{c}_{\theta_0,0}(u) \leq \varepsilon(\log n - \mathrm{KL}(\hat{\pi}_{\theta_0,\varepsilon}(u) \,||\, \hat{\mu} \otimes \hat{\nu})),$$

where $\hat{\pi}_{\theta_0,\varepsilon}(u)$ is the ε-regularized OT distribution with the cost function $u'\phi(\cdot,\cdot,\theta_0)$. Hence the (potentially conservative) test can be based on the adjusted statistic

$$\sqrt{n} \max_{u \in \mathbb{B}} \left(\hat{c}_{\theta_0,\varepsilon}(u) - \varepsilon(\log n - \mathrm{KL}(\hat{\pi}_{\theta_0,\varepsilon}(u) \,||\, \hat{\mu} \otimes \hat{\nu})) \right).$$

Remark 5. *Taking the minimum of $\sqrt{n}\hat{D}(\theta)$ over the parameter space $\theta \in \Theta$ yields a statistic that*

can be used to develop a specification test, i.e., a test of the hypothesis $\Theta_I \neq \varnothing$, see, e.g., Bugni et al. (2015) for a similar idea in the context of moment inequality models. The bootstrap of Fang and Santos (2019) can then be employed to obtain critical values for such a test.

3.3 Numerical implementation

Our procedure requires computing the distance statistic (3), which amounts to solving two nested optimization problems.

The inner problem (entropic OT) is solvable by a very fast and simple numerical procedure called the *Sinkhorn algorithm* described in Algorithm 2.

Algorithm 2 Sinkhorn algorithm for entropic OT with cost $(x, y) \mapsto u'\phi(x, y, \theta)$

Require: Samples $x_1, \ldots, x_n$; $y_1, \ldots, y_m$; moment function $\phi(\cdot, \cdot, \theta)$; vector $u \in \mathbb{B}$; regularization parameter $\varepsilon > 0$

Ensure: Approximate value $\hat{c}_\theta(u)$

Construct cost matrix:

$C_{ij} \leftarrow u'\phi(x_i, y_j, \theta)$ for all $i = 1, \ldots, n$, $j = 1, \ldots, m$

Initialize kernel and scaling vectors:

$K \leftarrow \exp(-C/\varepsilon)$

$a \leftarrow \mathbf{1}_n, \quad b \leftarrow \mathbf{1}_m$

Iterate to enforce marginal constraints:

while convergence criterion not met **do**

  $a \leftarrow \frac{1/n}{Kb}$

  $b \leftarrow \frac{1/m}{K'a}$

end while

Compute entropic OT value:

$P \leftarrow \operatorname{diag}(a)\, K\, \operatorname{diag}(b)$

$\hat{c}_\theta(u) \leftarrow \sum_{i=1}^n \sum_{j=1}^m P_{ij} C_{ij} \; + \; \varepsilon \sum_{i=1}^n \sum_{j=1}^m P_{ij}(\log P_{ij} - 1)$

return $\hat{c}_\theta(u)$

The outer problem is the maximization of a concave function $\hat{c}_\theta(u)$ (the output of Algorithm 2) over the unit ball $u \in \mathbb{B}$. We solve this problem using *projected gradient ascent* described in Algorithm 3. Of course, many other off-the-shelf algorithms are available for constrained concave optimization; for a comprehensive review, see Bubeck et al. (2015).

4 Illustrative examples

This section presents four empirical examples that illustrate how our methods can address partial identification problems arising from missing or incomplete data. The examples span different areas of econometrics: panel data, causal inference, and macro-finance. In each case, we show how the

Algorithm 3 Projected gradient ascent

Require: Initial point $u_0 \in \mathbb{B}$, step size $\eta > 0$, tolerance $\delta > 0$

Ensure: Approximate maximizer of $\hat{c}_\theta(u)$ over $u \in \mathbb{B}$

$t \leftarrow 0$

while $\left\|\nabla \hat{c}_\theta(u_t)\right\|_2 > \delta$ **do**

(1) Gradient step:

$u_{t+1} \leftarrow u_t + \eta \nabla \hat{c}_\theta(u_t)$

(2) Projection step:

$u_{t+1} \leftarrow u_{t+1} / \max\{1, \|u_{t+1}\|_2\}$

$t \leftarrow t + 1$

end while

return u_t

inability to observe certain joint distributions leads to partial identification of parameters of interest and how our methodology can be applied to characterize the identified sets in various contexts.

4.1 Fixed effects panel logit with attrition and refreshment

This is our leading example and forms the basis for the Monte Carlo simulations in Section 5. Panel logit models are widely used in empirical studies to analyze binary outcomes while controlling for individual heterogeneity through fixed effects. However, attrition is a common issue in panel data, where units may drop out of the sample in subsequent periods for reasons potentially correlated with outcomes of interest. When a refreshment sample is available, the model can be partially identified via OT. With an adjustment that exploits the panel structure, our methodology can be used to derive tight bounds for the common slope parameter and the AME.

4.1.1 Common slope parameter

The common slope parameter is point identified under complete data or when attrition is independent of outcomes conditional on observables, but becomes partially identified under unrestricted attrition. For notation simplicity, consider a static panel logit model with two periods $T = 2$,

$$Y_{it} = \mathbf{1}\left\{X_{it}'\theta + \alpha_i - \varepsilon_{it} > 0\right\}, \tag{4}$$

where θ captures the effect of covariates on the outcome, α_i represents individual fixed effects, and ε_{it} follows a standard logistic distribution.

The standard approach to eliminate the incidental parameter α_i is to condition on the sufficient statistic $S_i = Y_{i1} + Y_{i2}$. For individuals with $S_i = 1$ (switchers), the conditional log-likelihood is

$$\ell_i(\theta \mid S_i = 1) = Y_{i1}X_{i1}'\theta + Y_{i2}X_{i2}'\theta - \ln\left(e^{X_{i1}'\theta} + e^{X_{i2}'\theta}\right),$$

with the corresponding conditional score function $s(Y_{i1}, Y_{i2}, X_{i1}, X_{i2}; \theta)$ and the moment condition

$$\mathbb{E}\left[s(Y_{i1}, Y_{i2}, X_{i1}, X_{i2}; \theta_0) \mid S_i = 1\right] = 0.$$

The explicit form of the score is given in Appendix B.1. To translate this moment condition into our OT framework, we embed the event $\{Y_{i1} + Y_{i2} = 1\}$ into the cost function and define

$$\phi\left(y_1, y_2, x_1, x_2; \theta\right) = s(y_1, y_2, x_1, x_2; \theta)\mathbf{1}\left\{y_1 + y_2 = 1\right\}.$$

A naive approach to partially identify θ would use only the marginal distributions from period 1 (original sample) and period 2 (refreshment sample). However, we can achieve tighter bounds by exploiting the panel structure. Since we observe both periods for retainers, we fix their joint distribution $f_{1,2|\text{ret}}$ and apply OT only to couple the attriter distributions $f_{1|\text{att}}$ and $f_{2|\text{att}}$ across periods. While $f_{1|\text{att}}$ is directly observed, $f_{2|\text{att}}$ is unobserved because attriters are missing in period 2, but it can be recovered from the observed refreshment and retainer samples via the law of total probability. See Appendix B.1 for details.

Let p denote the retention rate. The attriter contribution to the moment bounds is

$$\underline{\nu}_{\text{att}}(\theta) = \inf_{f \in \Pi(f_{1|\text{att}}, f_{2|\text{att}})} \mathbb{E}_f[\phi(Y_1, Y_2, X_1, X_2; \theta)], \quad \overline{\nu}_{\text{att}}(\theta) = \sup_{f \in \Pi(f_{1|\text{att}}, f_{2|\text{att}})} \mathbb{E}_f[\phi(Y_1, Y_2, X_1, X_2; \theta)].$$

Since $\mathbb{E}_{f_{1,2|\text{ret}}}[\phi(Y_1, Y_2, X_1, X_2; \theta)]$ can be computed directly from the observed data for retainers, the overall bounds are

$$\begin{aligned}
\underline{\nu}(\theta) &= p \cdot \mathbb{E}_{f_{1,2|\text{ret}}}[\phi(Y_1, Y_2, X_1, X_2; \theta)] + (1-p) \cdot \underline{\nu}_{\text{att}}(\theta), \\
\overline{\nu}(\theta) &= p \cdot \mathbb{E}_{f_{1,2|\text{ret}}}[\phi(Y_1, Y_2, X_1, X_2; \theta)] + (1-p) \cdot \overline{\nu}_{\text{att}}(\theta).
\end{aligned}$$

The identified set for θ is therefore $\Theta_{I,0} = \{\theta : \underline{\nu}(\theta) \leq 0 \leq \overline{\nu}(\theta)\}$. Algorithm 4 in Appendix B.1 presents our estimation procedure.

Our estimator naturally extends to dynamic panel logit models where lagged dependent variables appear as regressors. It is possible to incorporate the moment conditions developed by Honoré and Weidner (2024) as the cost function, with a similar but more complicated partition structure separating retainers and attriters to achieve efficiency gains.

4.1.2 AME

We now consider estimation of the AME, which captures the average change in outcome probability induced by a marginal change in a covariate and provides a directly interpretable measure for

empirical work. Specifically, the AME of covariate j at period τ is defined as

$$\delta_{\tau,j} = \theta_j \, \mathbb{E}[\Lambda(X_\tau'\theta + \alpha)(1 - \Lambda(X_\tau'\theta + \alpha))].$$

The AME is partially identified even without attrition due to the incidental parameters problem combined with the nonlinear structure of the logit model. Under unrestricted attrition, this identification issue becomes more severe as the joint distribution of outcomes across periods is no longer observable.

Davezies et al. (2024) show how to construct outer bounds on the AME without attrition. They show that $\delta_{\tau,j}$ belongs to the interval $\tilde{\delta} \pm \bar{b}$, where $\tilde{\delta} = \mathbb{E}[p(X_{1:T}, S, \theta_0)]$ and $\bar{b} = \mathbb{E}[a(X_{1:T}, S, \theta_0)]$ for functions p and a constructed from a degree-$(T+1)$ Chebyshev polynomial. The explicit formulas for p, a, and the associated coefficients are collected in Section B.2.1, including closed-form expressions for the case $T = 2$.

Under unrestricted attrition, we extend the partition approach for θ to bound the AME. First, we compute the identified set $\widehat{\Theta}_I$ for the common slope parameters θ as in Algorithm 4 and construct a finite grid $\{\theta^{(g)}\} \subset \widehat{\Theta}_I$. Second, for each grid point $\theta^{(g)}$ we plug it into the Chebyshev approximation and, using our partition of retainers and attriters, compute the corresponding AME bounds $\left[\underline{\delta}(\theta^{(g)}), \overline{\delta}(\theta^{(g)})\right]$ via OT. Finally, we profile over θ by taking the union to obtain the identified set for the AME: $\bigcup_g \left[\underline{\delta}(\theta^{(g)}), \overline{\delta}(\theta^{(g)})\right]$. See Section B.2.2 for details.

Although this grid-based profiling yields conservative AME bounds due to the Chebyshev polynomial approximation and the two-stage approach that first estimates bounds on θ and then bounds on δ, it delivers a significant improvement in computational efficiency over alternative methods such as Hankel moment matrix positivity bounds or a single-stage approach that estimates both θ and δ simultaneously.

4.2 Nonlinear treatment effects

In causal inference, researchers face the fundamental problem of missing data because they observe either the potential outcome under treatment $Y(1)$ or under control $Y(0)$ for each individual, but never both simultaneously. This example demonstrates how OT methods provide identified sets when the object of interest is a functional of the joint distribution of potential outcomes. See also for example Fan et al. (2025).

Consider a binary treatment $D \in \{0, 1\}$, outcome Y, and pre-treatment covariates X. Let $Y(d)$ denote the potential outcome under treatment status d, with the observed outcome being $Y = DY(1) + (1 - D)Y(0)$. We impose the standard assumptions $(Y(0), Y(1)) \perp D \mid X$ (strong ignorability) and $0 < e(X) < 1$ almost surely (monotonicity), where $e(X) = P(D = 1 \mid X)$ is the propensity score.

The parameter of interest is

$$\theta = \mathbb{E}\left[h(Y(1), Y(0))\right],$$

where $h : \mathbb{R}^2 \to \mathbb{R}$ is a known function of both potential outcomes. This encompasses parameters such as the average treatment effect (ATE) $h(y_1, y_0) = y_1 - y_0$, distributional treatment effects $h(y_1, y_0) = \mathbf{1}\{y_1 - y_0 \leq t\}$, the variance of treatment effects $h(y_1, y_0) = (y_1 - y_0)^2$, and the share of units that benefit from treatment $h(y_1, y_0) = \mathbf{1}\{y_1 > y_0\}$. While the ATE is point identified, parameters involving the joint distribution of $(Y(1), Y(0))$ are generally only partially identified since we never observe both potential outcomes for the same individual.

To implement our OT approach, we employ propensity score stratification. By the propensity score theorem of Rosenbaum and Rubin (1983), we have $(Y(0), Y(1)) \perp D \mid e(X)$. Rather than conditioning on the covariate vector X directly, we discretize the propensity score into K strata, $\{I_k\}_{k=1}^K$, with equal probability $\pi > 0$ for each stratum. The propensity score stratification reduces the curse of dimensionality when X is high-dimensional, ensures sufficient sample sizes, and facilitates the implementation of OT estimation within each stratum.

Within each stratum k, the conditional distributions $F_{Y(d)|e(X)\in I_k}(y)$ for $d \in \{0, 1\}$ are point identified. However, the joint distribution $F_{Y(1),Y(0)|e(X)\in I_k}(y_1, y_0)$ remains unknown, leading to partial identification of stratum-specific parameters $\theta_k = \mathbb{E}[h(Y(1), Y(0)) \mid e(X) \in I_k]$. The identified set for θ_k is given by the OT bounds $\Theta_k = \left[\underline{\theta}_k, \overline{\theta}_k\right]$, where the cost function is $\phi(y_1, y_0) = h(y_1, y_0)$,

$$\underline{\theta}_k = \inf_{F \in \Pi\left(F_{1|k}, F_{0|k}\right)} \mathbb{E}_F[\phi(Y(1), Y(0)) \mid e(X) \in I_k],$$
$$\overline{\theta}_k = \sup_{F \in \Pi\left(F_{1|k}, F_{0|k}\right)} \mathbb{E}_F[\phi(Y(1), Y(0)) \mid e(X) \in I_k],$$

and $F_{d|k} = F_{Y(d)|e(X)\in I_k}$ denotes the marginal distribution of $Y(d)$ in stratum k. The parameter θ is then bounded by $\theta \in \pi \left[\sum_{k=1}^K \underline{\theta}_k, \sum_{k=1}^K \overline{\theta}_k\right]$.

The bounds derived above are sharp for many functions h of interest. As shown in Fan et al. (2017), when h is either a supermodular functional or an indicator of a nondecreasing transformation, the sharp bounds are given by the Fréchet-Hoeffding bounds, which are achieved by comonotone and counter-comonotone couplings (corresponding to perfect positive and negative dependence, respectively).

4.3 Nonparametric IV without large support

IV methods are commonly used in causal inference when strong ignorability fails. The classical control function approach to IV requires a large support condition for point identification. However, this assumption often fails in practice when the treatment is discrete or has limited variation, such

as binary treatments in Angrist et al. (1996) and discrete treatment intensity in Angrist and Imbens (1995). This example shows how our methodology can deliver meaningful bounds even when the large support assumption fails.

Consider the standard nonparametric IV model

$$Y = g(X, V), \quad X = h(Z, W), \quad (W, V) \perp Z,$$

where Y is the outcome, X is the endogenous treatment, Z is the instrument, and (V, W) are the unobserved errors. We want to estimate a generic object of interest $\theta = \mathbb{E}[\Lambda(g(X, V))]$ beyond the local average treatment effects (LATE), e.g., $\Lambda(y) = y$ for the ATE, or $\Lambda(y) = \mathbf{1}\{y \leq t\}$ for the distributional treatment effect. The classical control variable formula requires not only treatment monotonicity ($w \mapsto h(z, w)$ strictly increasing) but also large support ($\text{supp}(R \mid V = v) = \text{supp}(R)$ for all v, where $R = F_{X|Z}(X)$), see, e.g., Section 4.1 in Gunsilius (2025). When this large support condition fails, θ becomes partially identified.

To formalize this setting, we impose two assumptions. First, similar to the classical setup, we assume treatment monotonicity: $w \mapsto h(z, w)$ is strictly increasing for each z, so under a monotone transformation, we can define the control variable $W = F_{X|Z}(X)$. Second, we also assume outcome monotonicity: $v \mapsto g(x, v)$ is strictly increasing, so define $V = F_{Y|X}(Y) \sim \text{Uniform}[0, 1]$. Note that W and V are, in general, not independent. If W has limited variation, such as discrete or censored treatment, the large support condition may fail, i.e., $\text{supp}(W \mid V = v) \subsetneq \text{supp}(W)$ for some v.

In many empirical applications, the instrument Z is supported on a finite number of values. For example, Angrist and Imbens (1995) use quarter-of-birth dummies as instruments for schooling, and Card (1995) uses a binary instrument based on proximity to four-year colleges. For each z, the marginals F_W and F_V can be recovered from the data, and the moment function is

$$\phi(w, v; z) = \Lambda\left(g\left(F^{-1}_{X|Z=z}(w), v\right)\right).$$

Then the sharp identified set for θ is the interval $\left[\underline{\theta}, \overline{\theta}\right]$ with

$$\underline{\theta} = \inf_{F \in \Pi(F_W, F_V)} \sum_z \mathbb{E}_F[\phi(W, V; z)] \Pr(Z = z), \quad \bar{\theta} = \sup_{F \in \Pi(F_W, F_V)} \sum_z \mathbb{E}_F[\phi(W, V; z)] \Pr(Z = z).$$

4.4 Euler equation estimation with repeated cross-sections

A prominent example in macro-finance is estimating the discount factor β and risk aversion γ from the constant relative risk aversion (CRRA) Euler equation

$$\mathbb{E}\left[\beta \left(C_{i,t+1}/C_{it}\right)^{-\gamma} R_{t+1} - 1 \mid \mathcal{I}_{it}\right] = 0,$$

where C_{it} is individual i's consumption at time t, R_{t+1} is the common asset return, and $\mathcal{I}_{it}$ is the information set.

In practice, this single nonlinear conditional moment is converted into an overidentified unconditional GMM by introducing a k-dimensional vector of instruments with $k \geq 2$, $Z_{it} = \left(Z_t^A, Z_{it}^I\right)$, where Z_t^A are lagged macro variables (such as GDP growth rates and interest rates), and Z_{it}^I are lagged individual variables (such as demographics, as well as prior income and consumption). The moment condition then becomes

$$\mathbb{E}\left[Z_{it}\left(\beta\left(C_{i,t+1}/C_{it}\right)^{-\gamma} R_{t+1} - 1\right)\right] = 0.$$

Under a standard rank condition, this equation can be used to estimate (β, γ).

To estimate the Euler equation, one would ideally use data that preserve cross-sectional heterogeneity while providing a sufficient sample size, which in practice motivates the use of repeated cross-sections or short rotating panels. Much of the early empirical literature, however, relied on aggregate time-series consumption and return data, as in Hansen and Singleton (1982). Because the Euler equation is nonlinear in consumption, Jensen's inequality implies that the nonlinear moment evaluated at aggregate consumption would differ from the cross-sectional average of individual-level terms. This mismatch can induce systematic bias in GMM estimates. Subsequent work emphasized granular data to retain heterogeneity. For example, Dynan et al. (2004) exploit the panel structure of the Panel Study of Income Dynamics (PSID), but the PSID has a relatively small cross-sectional sample size. In contrast, many large-scale household surveys, such as the Consumer Expenditure Survey, offer rich cross-sectional coverage via repeated cross-sections or short rotating panels, making them well-suited for our OT-based approach. See also Liu and Plagborg-Møller (2023) on estimating full structural models with repeated cross-sections.

In an extreme case, suppose we observe only repeated cross-sections, so that the marginal distributions $f_{C_{it}, Z_{it}^I}(c, z^I)$ and $f_{C_{i,t+1}}(\tilde{c})$, are known, in contrast to their joint law. Let $\theta = (\beta, \gamma)'$. Then the parameter is only partially identified with the identified set $\Theta = \{\theta : \underline{\nu}(\theta) \leq 0 \leq \overline{\nu}(\theta)\}$, where

$$\underline{\nu}(\theta) = \inf_{f \in \Pi\left(f_{C_{it}, Z_{it}^I}, f_{C_{i,t+1}}\right)} \mathbb{E}_f\left[\phi(Z_{it}, C_{it}, C_{i,t+1}, R_{t+1}; \theta)\right],$$

and $\overline{\nu}(\theta)$ is the supremum of the same expression. Here $\Pi(\cdot)$ is the set of all couplings consistent with the observed marginals, and the moment function is

$$\phi(z, c, \tilde{c}, r; \theta) = z'\left(\beta\left(\tilde{c}/c\right)^{-\gamma} r - 1\right).$$

With short rotating panels, where households are observed for only several periods, the OT problem

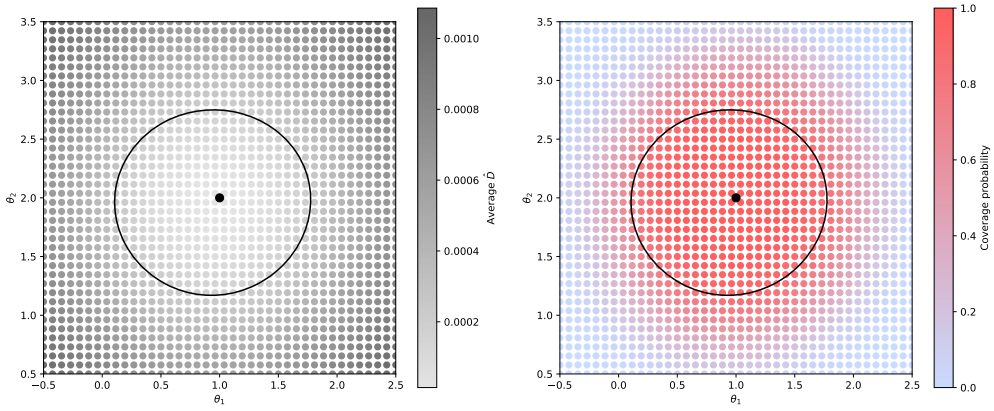

Figure 2: Left panel: distance $\hat{D}(\theta)$; right panel: coverage probability of the 90% confidence region. Fixed effects logit with attrition and refreshment.

becomes more complex: one can exploit the limited longitudinal links to tighten the identified set while using OT for the remaining unlinked portions of the data.

5 Monte Carlo simulation

In this section, we give a small illustration of the performance of our inference procedure.

The data-generating process is the fixed effects panel logit with attrition and refreshment as described in Section 4.1 and defined in (4). The true parameter is $\theta_0 = (1.0, 2.0)'$. The covariate vector $X_{it} = (X_{it,1}, X_{it,2})'$ consists of two independent components $X_{it,1}$ and $X_{it,2}$ that have discrete uniform distributions on three-valued sets $\mathcal{X}_1 = \{0.42, 0.55, 0.60\}$ and $\mathcal{X}_2 = \{0.54, 0.65, 0.72\}$, respectively, and are independent across units i and time t. The fixed effects α_i are drawn i.i.d. from the standard normal distribution. The idiosyncratic error terms ε_{it} are drawn i.i.d. from the standard logistic distribution with zero mean and unit scale. The sizes of both the first-period and the refreshment samples are $n_{\text{org}} = n_{\text{ref}} = 15000$. The attrition rate is $1 - p = 10\%$, and the units drop out of the sample completely at random.

We conduct 400 simulations. For each simulation, we construct the 90% confidence region by inverting the test in Algorithm 1 on a grid of hypothesized values $\theta^* \in [-0.5, 2.5] \times [0.5, 3.5]$ centered at the true value $\theta_0 = (1.0, 2.0)'$. We use Algorithm 2 to solve the discrete entropic OT problem on a grid, where each of the two marginals is defined on $2 \times 3 \times 3$ values in the joint support $\{0, 1\} \times \mathcal{X}_1 \times \mathcal{X}_2$ of $(y_{it}, x_{it1}, x_{it2})$. We then use grid search on the unit circle to calculate the distance $\hat{D}(\theta^*)$. We set the entropic regularization parameter $\varepsilon = 0.1$ and the tuning parameter $\iota_n = 0.05 n^{-1/2} \log n$ for constructing the enlarged argmax set in the bootstrap procedure.

Figure 2 shows the simulation results. Each value θ^* in the grid is colored according to the value

of $\hat{D}(\theta^*)$ (left panel) or the proportion of times it is covered by the 90% confidence region (right panel). The identified set $\Theta_{I,\varepsilon}$ is indicated by the black contour line. We see that the distance statistic tracks the identified set and the confidence region performs reasonably well.

6 Conclusion

This paper develops a methodology for estimation and inference in GMM where the distribution of the data is identified only up to its marginals. We characterize the identified set for the parameter of interest using tools from convex analysis and OT. The practical implementation of classical OT is hindered by both theoretical and computational limitations. To overcome these issues, we rely on the regularized (entropic) version of OT. The resulting OT-based characterization directly informs an estimator and a test statistic for conducting inference and constructing confidence regions.

We establish a central limit theorem for the entropic OT value under smooth cost functions and use it to show $\sqrt{n}$-consistency and asymptotic normality of our proposed statistic. We then obtain valid critical values via the bootstrap for directionally differentiable functionals developed in Fang and Santos (2019).

Our estimation and inference methodology is generic and computationally efficient. It is also highly relevant for applied work, since many important economic questions, ranging from the effects of policy interventions to the dynamics of household behavior, involve parameters that are characterized via OT-based partially identified GMM.

Our framework admits several promising theoretical extensions. First, it naturally extends to settings with more than two marginals, such as panel data with multiple waves or repeated cross-sections over several periods (multi-marginal OT). Second, it would be of interest to extend it to GMM models with conditional moment restrictions. Finally, when the moment conditions underidentify the parameter even under point identification of the data distribution, the interaction between these two sources of partial identification is theoretically challenging and deserving of further study. We leave these extensions for future work.

References

Aliprantis, C. D. and K. C. Border (2006): *Infinite dimensional analysis: a hitchhiker's guide*, Springer.

Andrews, D. W. (2000): "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space," *Econometrica*, 399–405.

Andrews, D. W. and P. Guggenberger (2009): "Validity of subsampling and "plug-in asymp-

totic" inference for parameters defined by moment inequalities," *Econometric Theory*, 25, 669–709.

Andrews, D. W. and S. Han (2009): "Invalidity of the bootstrap and the m out of n bootstrap for confidence interval endpoints defined by moment inequalities," *The Econometrics Journal*, 12, S172–S199.

Andrews, D. W. and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666.

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90, 431–442.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Armstrong, T. B. (2015): "Asymptotically exact inference in conditional moment inequality models," *Journal of Econometrics*, 186, 51–65.

——— (2018): "On the choice of test statistic for conditional moment inequalities," *Journal of Econometrics*, 203, 241–255.

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

Beresteanu, A. and F. Molinari (2008): "Asymptotic properties for a class of partially identified models," *Econometrica*, 76, 763–814.

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set identified linear models," *Econometrica*, 80, 1129–1155.

Botosaru, I., I. Loh, and C. Muris (2024): "An Adversarial Approach to Identification," *arXiv preprint arXiv:2411.04239*.

Bubeck, S. et al. (2015): "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, 8, 231–357.

Bugni, F. A. (2010): "Bootstrap inference in partially identified models defined by moment inequalities: Coverage of the identified set," *Econometrica*, 78, 735–753.

Bugni, F. A., I. A. Canay, and X. Shi (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185, 259–282.

Canay, I. A. and A. M. Shaikh (2017): "Practical and theoretical advances in inference for par-

tially identified models," in *Advances in Economics and Econometrics: Eleventh World Congress*, Cambridge University Press, vol. 2, 271–306.

Card, D. (1995): "Using geographic variation in college proximity to estimate the returns to schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, University of Toronto Press, 201–222.

Chetverikov, D. (2018): "Adaptive tests of conditional moment inequalities," *Econometric Theory*, 34, 186–227.

Chiong, K. X., A. Galichon, and M. Shum (2016): "Duality in dynamic discrete-choice models," *Quantitative Economics*, 7, 83–115.

Constantine, G. and T. Savits (1996): "A multivariate Faa di Bruno formula with applications," *Transactions of the American Mathematical Society*, 348, 503–520.

Cuturi, M. (2013): "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, 26.

Davezies, L., X. D'Haultfoeuille, and L. Laage (2024): "Identification and estimation of average marginal effects in fixed effects logit models," *arXiv preprint arXiv:2105.00879*.

Dynan, K. E., J. Skinner, and S. P. Zeldes (2004): "Do the Rich Save More?" *Journal of Political Economy*, 112, 397–444.

Fan, Y., E. Guerre, and D. Zhu (2017): "Partial identification of functionals of the joint distribution of potential outcomes," *Journal of Econometrics*, 197, 42–59.

Fan, Y., B. Pass, and X. Shi (2025): "Partial Identification in Moment Models with Incomplete Data via Optimal Transport," *arXiv preprint arXiv:2503.16098*.

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

Firpo, S. and G. Ridder (2019): "Partial identification of the treatment effect distribution and its functionals," *Journal of Econometrics*, 213, 210–234.

Franguridi, G., J. Hahn, P. Hoonhout, A. Kapteyn, and G. Ridder (2025a): "Raking for estimation and inference in panel models with nonignorable attrition and refreshment," *arXiv preprint arXiv:2512.13270*.

Franguridi, G., J. Hahn, and G. Ridder (2025b): "Robust estimation and inference for panels with nonignorable attrition and refreshment," *Working paper*.

Franguridi, G. and H. R. Moon (2025): "Generalized method of moments with partially missing data," *arXiv preprint arXiv:2511.21988*.

Galichon, A. and B. Salanié (2022): "Cupid's invisible hand: Social surplus and identification in matching models," *The Review of Economic Studies*, 89, 2600–2629.

Goldfeld, Z., K. Kato, G. Rioux, and R. Sadhu (2024): "Statistical inference with regularized optimal transport," *Information and Inference: A Journal of the IMA*, 13, iaad056.

Gunsilius, F. and Y. Xu (2021): "Matching for causal effects via multimarginal unbalanced optimal transport," *arXiv preprint arXiv:2112.04398*.

Gunsilius, F. F. (2025): "A primer on optimal transport for causal inference with observational data," *arXiv preprint arXiv:2503.07811*.

Hansen, L. P. and K. J. Singleton (1982): "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, 50, 1269–1286.

Hazard, Y. and T. Kitagawa (2025): "Who With Whom? Learning Optimal Matching Policies," *arXiv preprint arXiv:2507.13567*.

Honoré, B. E. and M. Weidner (2024): "Moment conditions for dynamic panel logit models with fixed effects," *Review of Economic Studies*, rdae097.

Hundrieser, S., M. Klatt, A. Munk, and T. Staudt (2024): "A unifying approach to distributional limits for empirical optimal transport," *Bernoulli*, 30, 2846–2877.

Imbens, G. W. and C. F. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72, 1845–1857.

Liu, L. and M. Plagborg-Møller (2023): "Full-information estimation of heterogeneous agent models using macro and micro data," *Quantitative Economics*, 14, 1–35.

Makarov, G. (1982): "Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed," *Theory of Probability & its Applications*, 26, 803–806.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 80, 319–323.

——— (2003): *Partial Identification of Economic Models*, Cambridge University Press.

Meango, R., M. Henry, and I. Mourifie (2025): "Combining stated and revealed preferences," *arXiv preprint arXiv:2507.13552*.

Mena, G. and J. Niles-Weed (2019): "Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem," *Advances in neural information processing systems*, 32.

Pooladian, A.-A., M. Cuturi, and J. Niles-Weed (2022): "Debiaser beware: Pitfalls of

centering regularized transport maps," in *International conference on machine learning*, PMLR, 17830–17847.

Romano, J. P. and A. M. Shaikh (2010): "Inference for the identified set in partially identified econometric models," *Econometrica*, 78, 169–211.

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Schennach, S. and V. Starck (2025): "Optimally-Transported Generalized Method of Moments," *arXiv preprint arXiv:2511.05712*.

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

Stoye, J. (2009): "More on confidence intervals for partially identified parameters," *Econometrica*, 77, 1299–1315.

Tamer, E. (2010): "Partial identification in econometrics," *Annu. Rev. Econ.*, 2, 167–195.

Torous, W., F. Gunsilius, and P. Rigollet (2024): "An optimal transport approach to estimating causal effects via nonlinear difference-in-differences," *Journal of Causal Inference*, 12.

van der Vaart, A. and J. Wellner (2023): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer.

Van der Vaart, A. W. (2000): *Asymptotic statistics*, vol. 3, Cambridge university press.

Voronin, A. (2025): "Generalized Optimal Transport," *arXiv preprint arXiv:2507.22422*.

Appendices

A Proofs of theoretical results

A.1 Proof of Theorem 1

We follow the proof of Proposition 1 in Franguridi and Moon (2025) closely. By Assumption 1(iii), $\Theta_{I,0}$ is nonempty.

Step 1: $\nu_\Pi(\theta)$ is compact and convex.

Convexity is trivial. In view of Assumptions 1(ii) and 1(iv), Theorem 15.11 and Corollary 15.7 in Aliprantis and Border (2006) imply that Π is compact in the weak topology and the map $\pi \mapsto \mathbb{E}_\pi \phi(X, Y, \theta)$ is continuous. Hence, $\nu_\Pi(\theta)$ is compact as the image of a compact set under a continuous map.

Step 2: $\Theta_{I,0} = D_0^{-1}(\{0\})$.

Since $\nu_\Pi(\theta)$ is closed and convex, $0 \in \nu_\Pi(\theta)$ if and only if its support function is everywhere nonnegative

$$\psi_{\nu_\Pi(\theta)}(u) := \max_{\pi \in \Pi} u' \, \mathbb{E}_\pi \, \phi(X, Y, \theta) \geq 0 \text{ for all } u \in \mathbb{R}^{\dim(\phi)}.$$

Indeed, if $0 \in \nu_\Pi(\theta)$, then $\psi_{\nu_\Pi(\theta)}(u) = \max_{\nu \in \nu_\Pi(\theta)} u'\nu \geq 0$. Conversely, if $0 \notin \nu_\Pi(\theta)$, then by strong separation of a point from the closed convex set $\nu_\Pi(\theta)$, there exists $u \neq 0$ and $\alpha > 0$ such that $u'\nu \leq \alpha$ for all $\nu \in \nu_\Pi(\theta)$. This implies $\psi_{\nu_\Pi(\theta)}(u) \leq -\alpha < 0$.

Since $\psi_{\nu_\Pi(\theta)}(0) = 0$ for all θ, nonnegativity of $\psi_{\nu_\Pi(\theta)}$ is equivalent to the equality of

$$\min_{u \in \mathbb{R}^{\dim(\phi)}} \psi_{\nu_\Pi(\theta)}(u) \tag{5}$$

to zero. Let us show that the latter is equivalent to the equality of

$$\min_{u \in \mathbb{B}} \psi_{\nu_\Pi(\theta)}(u) \tag{6}$$

to zero. Indeed, if (5) is zero, then the minimum is achieved at $u = 0$, and hence (6) is zero. Conversely, if (5) is nonzero, then there exists u such that $\psi_{\nu_\Pi(\theta)}(u) < 0$. By the positive homogeneity of the support function, $\psi_{\nu_\Pi(\theta)}(u/\|u\|) = \psi_{\nu_\Pi(\theta)}(u)/\|u\| < 0$, and hence the expression (6) is negative.

Step 3: $\Theta_{I,0}$ is compact.

Since $\Theta_{I,0} = D^{-1}(\{0\})$, it suffices to establish the continuity of $D(\theta)$. For this, notice that

$$(u, \theta, \pi) \mapsto u' \, \mathbb{E}_\pi \, \phi(X, Y, \theta)$$

is a continuous function on the compact set $\mathbb{B} \times \Theta \times \Pi$, where Π is equipped with the weak topology. Berge's maximum theorem (see, e.g., Theorem 17.31 in Aliprantis and Border (2006)) implies that the function

$$(u, \theta) \mapsto \max_{\pi \in \Pi} u' \, \mathbb{E}_\pi \, \phi(X, Y, \theta) = \psi_{\nu_\Pi(\theta)}(u)$$

is continuous on the compact set $\mathbb{B} \times \Theta$. Applying Berge's theorem again implies that $D(\theta)$ is continuous, completing the proof.

A.2 Proof of Theorem 2

By Assumption 2(i), the set $\Theta_I = \{\theta \in \Theta : D(\theta) = 0\}$ is nonempty. Moreover, by Assumption 1 and the same Berge-type argument as in the proof of Theorem 1, the map $\theta \mapsto D(\theta)$ is continuous on the compact set Θ. Hence Θ_I is a compact subset of Θ.

We will show that $\Theta_I \subset \widehat{\Theta}_I$ w.p.a. 1 and that for every $\delta > 0$, $\widehat{\Theta}_I \subset \Theta_I^\delta$ w.p.a. 1, where Θ_I^δ is the δ-enlargement of Θ_I, i.e.

$$\Theta_I^\delta = \{\theta \in \Theta : \, d(\theta, \Theta_I) \leq \delta\}$$

Nonemptiness and compactness of Θ_I will then imply the required Hausdorff convergence.

First, let us show that $\Theta_I \subset \widehat{\Theta}_I$ w.p.a. 1. We have

$$\sup_{\theta \in \Theta_I} \hat{D}(\theta) \leq \sup_{\theta \in \Theta_I} D(\theta) + \|\hat{D} - D\|_\infty = \|\hat{D} - D\|_\infty.$$

By Assumption 2(ii) and Assumption 2(iii), the right-hand side is smaller than η_n w.p.a. 1, and hence $\Theta_I \subset \widehat{\Theta}_I$ w.p.a. 1.

Now, let us show that for every $\delta > 0$, $\widehat{\Theta}_I \subset \Theta_I^\delta$ w.p.a. 1. By Assumption 2(i), we have $\inf_{\theta \notin \Theta_I^\delta} D(\theta) \geq m(\delta) > 0$. On the event $\|\hat{D} - D\|_\infty \leq r_n$, we have

$$\inf_{\theta \notin \Theta_I^\delta} \hat{D}(\theta) \geq \inf_{\theta \notin \Theta_I^\delta} D(\theta) - \|\hat{D} - D\|_\infty \geq m(\delta) - r_n.$$

Choose n large enough so that $r_n \leq \eta_n < m(\delta)/2$. Then

$$\inf_{\theta \notin \Theta_I^\delta} \hat{D}(\theta) \geq m(\delta)/2 > \eta_n,$$

and hence no θ outside of Θ_I^δ belongs to $\widehat{\Theta}_I$. Therefore, $\widehat{\Theta}_I \subset \Theta_I^\delta$ w.p.a. 1.

A.3 Proof of Theorem 3

For a function $f \in C^s(\mathcal{X} \times \mathcal{Y})$, denote its Hölder norm by

$$\|f\|_{C^s(\mathcal{X}\times\mathcal{Y})} = \max_{0\le|\alpha|\le s} \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |\nabla^\alpha f(x,y)|\,,$$

where the maximum is taken over all multi-indices $\alpha = (\alpha_1, \ldots, \alpha_{2d}) \in \mathbb{N}_0^{2d}$ or order $|\alpha| := \alpha_1 + \cdots + \alpha_{2d} \le s$, and the partial derivative operator

$$\nabla^\alpha f(x,y) := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}\partial y_1^{\alpha_{d+1}}\cdots\partial y_d^{\alpha_{2d}}} f(x,y).$$

Assumptions 1(i), 3(i), and 3(ii) imply that there exist finite constants $\bar{\phi}$ and $\bar{\phi}_s$ such that

$$\sup_{u\in\mathbb{B}}\sup_{\theta\in\Theta}\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} |u'\phi(x,y,\theta)| \le \bar{\phi}$$

and

$$\sup_{u\in\mathbb{B}}\sup_{\theta\in\Theta} \|u'\phi(\cdot,\cdot,\theta)\|_{C^s(\mathcal{X}\times\mathcal{Y})} \le \bar{\phi}_s.$$

Define the mapping $c : \ell^\infty(\mathscr{F}) \to \ell^\infty(\mathbb{B}\times\Theta)$ by

$$c(\mu)(u,\theta) = \sup_{\varphi,\psi} \int \varphi\, d\mu + \int \psi\, d\nu - \int e^{\varphi\oplus\psi - u'\phi(\cdot,\cdot,\theta)}\, d\mu\otimes\nu + 1$$

Proof. **Part (i)**. Since $s > d/2$, the class

$$\mathscr{F} = \{f \in C^s(\mathcal{X}) \text{ such that } \|f\|_{C^s} \le C_{s,d}\}\,,$$

where $C_{s,d}$ is the constant in Proposition 2, is Donsker, see, e.g., Corollary 2.7.2 in van der Vaart and Wellner (2023). Therefore,

$$\sqrt{n}(\hat{\mu} - \mu) \rightsquigarrow \mathbb{G} \text{ in } \ell^\infty(\mathscr{F}),$$

where $\mathbb{G}$ is the generalized μ-Brownian bridge. In view of Proposition 3, we can apply the delta method for random measures in Proposition 1 of Goldfeld et al. (2024) to obtain

$$\sqrt{n}(\theta(\hat{\mu}) - \theta(\mu)) \rightsquigarrow \theta'_\mu(\mathbb{G}) = \mathbb{G}(\varphi) \sim N(0, \mathbb{V}_\mu(\varphi)).$$

Since θ'_μ is the point evaluation at φ, Proposition 2 in Goldfeld et al. (2024) implies that $\theta(\hat{\mu})$ is semiparametrically efficient.

Finally, since θ'_μ is linear, $\rho \mapsto \theta(\rho)$ is (fully) Hadamard differentiable at $\rho = \mu$ tangentially to $\operatorname{supp} \mathbb{G}$ by Corollary 1 in Goldfeld et al. (2024). Hence, nonparametric bootstrap is consistent by Theorem 23.9 in Van der Vaart (2000).

Part (ii). Define the function class

$$\mathscr{F}^{\oplus} = \{\varphi \oplus \psi : (\varphi, \psi) \text{ satisfies (10), (11)}\}.$$

Let us show that

$$|\theta(\mu_1, \nu_1) - \theta(\mu_0, \nu_0)| \le \|\mu_1 \otimes \nu_1 - \mu_0 \otimes \nu_0\|_{\mathscr{F}^{\oplus}} \tag{7}$$

for any probability measures μ_0, μ_1 on $\mathcal{X}$ and probability measures ν_0, ν_1 on $\mathcal{Y}$. Let φ_{ij}, ψ_{ij} be optimal potentials for (μ_i, ν_j) satisfying (10), (11). Then (15) implies

$$\begin{aligned}
\theta(\mu_1, \nu_1) - \theta(\mu_0, \nu_0) &= \theta(\mu_1, \nu_1) - \theta(\mu_0, \nu_1) + \theta(\mu_0, \nu_1) - \theta(\mu_0, \nu_0) \\
&\ge \int \varphi_{01}\, d(\mu_1 - \mu_0) + \int \psi_{00}\, d(\nu_1 - \nu_0) \\
&= \int (\varphi_{01} \oplus \psi_{00})\, d(\mu_1 \otimes \nu_1 - \mu_0 \otimes \nu_0).
\end{aligned}$$

Similarly, (16) implies

$$\theta(\mu_1, \nu_1) - \theta(\mu_0, \nu_0) \le \int (\varphi_{11} \oplus \psi_{01})\, d(\mu_1 \otimes \nu_1 - \mu_0 \otimes \nu_0).$$

Since $\varphi_{ij} \oplus \psi_{k\ell} \in \mathscr{F}^{\oplus}$, we obtain (7).

Moreover, arguing as in the proof of Proposition 3, we obtain

$$\lim_{t \downarrow 0} \frac{\theta(\mu_0 + t(\mu_1 - \mu_0), \mu_0 + t(\nu_1 - \nu_0)) - \theta(\mu_0, \nu_0)}{t} = \int (\varphi_{00} \oplus \psi_{00})\, d(\mu_1 \otimes \nu_1 - \mu_0 \otimes \nu_0).$$

Define $\mathscr{P}_0$ as the set of probability measures of the form $\rho_1 \otimes \rho_2$, where ρ_1, ρ_2 concentrate in $\mathcal{X}$, $\mathcal{Y}$, respectively. Applying Proposition 1 in Goldfeld et al. (2024) to the statement in Proposition 4 for $\delta(\rho_1 \otimes \rho_2) = \theta(\rho_1, \rho_2)$ and $\mathscr{F} = \mathscr{F}^{\oplus}$ yields

$$\begin{aligned}
\sqrt{n}(\theta(\hat{\mu}_n, \hat{\nu}_n) - \theta(\mu, \nu)) &= \sqrt{n}(\delta(\hat{\mu}_n \otimes \hat{\nu}_n) - \delta(\mu \otimes \nu)) \\
&\rightsquigarrow \delta'_{\mu\otimes\nu}(\mathbb{G}_{\mu\otimes\nu}) = \mathbb{G}_{\mu\otimes\nu}(\varphi \oplus \psi) \sim N(0, \mathbb{V}_\mu(\varphi) + \mathbb{V}_\nu(\psi)).
\end{aligned}$$

Finally, $\mathbb{V}_\mu(\varphi) + \mathbb{V}_\nu(\psi)$ is the semiparametric variance bound due to Corollary 2 of Goldfeld et al. (2024). □

The following result and its proof follow Proposition A.1 in Mena and Niles-Weed (2019).

Proposition 1. *For any $u \in \mathbb{B}$ and $\theta \in \Theta$, there exist optimal potentials $\varphi_{u,\theta}, \psi_{u,\theta}$ such that $\varphi_{u,\theta} \in C^s(\mathcal{X})$, $\psi \in C^s(\mathcal{Y})$, and*

$$|\varphi_{u,\theta}(x)| \leq \bar{\phi}, \tag{8}$$
$$|\psi_{u,\theta}(y)| \leq \bar{\phi}, \tag{9}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Proof. Fix $u \in \mathbb{B}$ and $\theta \in \Theta$ and let $(\varphi^0_{u,\theta}, \psi^0_{u,\theta})$ be any pair of optimal potentials. Assume without loss of generality that $u'\phi(\cdot,\cdot,\theta) \geq 0$. Since $(\varphi^0_{u,\theta} - a, \psi^0_{u,\theta} + a)$ is also a pair of optimal potentials for any $a \in \mathbb{R}$, we can assume

$$\int \varphi^0_{u,\theta}(x)\, d\mu(x) = \int \psi^0_{u,\theta}(y)\, d\nu(y) = \frac{1}{2} c(\mu,\nu)(u,\theta) \geq 0.$$

Define

$$\varphi_{u,\theta}(x) = -\log \int e^{\psi^0_{u,\theta}(y) - u'\phi(x,y,\theta)}\, d\nu(y),$$
$$\psi_{u,\theta}(y) = -\log \int e^{\varphi^0_{u,\theta}(x) - u'\phi(x,y,\theta)}\, d\mu(x).$$

Jensen's inequality combined with $\int \psi^0_{u,\theta}(y)\, d\nu(y) \geq 0$ yield

$$\varphi_{u,\theta}(x) \leq -\int (\psi^0_{u,\theta}(y) - u'\phi(x,y,\theta))\, d\nu(y) \leq \int u'\phi(x,y,\theta)\, d\nu(y) \leq \bar{\phi}.$$

To establish the lower bound on $\varphi_{u,\theta}(x)$, notice that, by Jensen's inequality,

$$\begin{aligned}
\psi^0_{u,\theta}(y) &= -\log \int e^{\varphi^0_{u,\theta}(x) - u'\phi(x,y,\theta)}\, d\mu(x) \\
&\leq -\int \varphi^0_{u,\theta}(x)\, d\mu(x) + \int u'\phi(x,y,\theta)\, d\mu(x) \\
&\leq \int u'\phi(x,y,\theta)\, d\mu(x).
\end{aligned}$$

Therefore,

$$\exp\{\psi^0_{u,\theta}(y) - u'\phi(x,y,\theta)\} \le \exp\left\{\int u'\phi(x,y,\theta)\,d\mu(x) - u'\phi(x,y,\theta)\right\} \le \exp\{\bar{\phi}\}.$$

Integrating this inequality w.r.t. ν and taking $-\log$ yields

$$\varphi_{u,\theta}(x) \ge -\bar{\phi}.$$

Notice that $\varphi_{u,\theta} \in C^s(\mathcal{X})$ and $\psi_{u,\theta} \in C^s(\mathcal{Y})$ by the dominated convergence theorem.

It remains to show that $\varphi_{u,\theta}, \psi_{u,\theta}$ are optimal potentials. By construction,

$$\int e^{\varphi_{u,\theta}(x)+\psi(y)-u'\phi(x,y,\theta)}\,d\mu(x) = 1 \text{ for all } y \in \mathcal{Y}.$$

Moreover,

$$\begin{aligned}\int e^{\varphi_{u,\theta}(x)+\psi_{u,\theta}(y)-u'\phi(x,y,\theta)}\,d\mu(x)d\nu(y) &= \int e^{\varphi_{u,\theta}(x)+\psi^0_{u,\theta}(y)-u'\phi(x,y,\theta)}\,d\mu(x)d\nu(y)\\ &= \int e^{\varphi^0_{u,\theta}(x)+\psi^0_{u,\theta}(y)-u'\phi(x,y,\theta)}\,d\mu(x)d\nu(y).\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}&\int (\varphi_{u,\theta}(x) - \varphi^0_{u,\theta}(x))\,d\mu(x) + \int (\psi_{u,\theta}(y) - \psi^0_{u,\theta}(y))\,d\nu(y)\\ &\ge -\log \int e^{\varphi_{u,\theta}(x)-\varphi^0_{u,\theta}(x)}\,d\mu(x) - \log \int e^{\psi_{u,\theta}(y)-\psi^0_{u,\theta}(y)}\,d\nu(y)\\ &= -\log \int e^{\varphi^0_{u,\theta}(x)+\psi^0_{u,\theta}(x)-u'\phi(x,y,\theta)}\,d\mu(x)d\nu(y) - \log \int e^{\varphi_{u,\theta}(x)+\psi^0_{u,\theta}(x)-u'\phi(x,y,\theta)}\,d\mu(x)d\nu(y)\\ &= 0.\end{aligned}$$

Since $(\varphi^0_{u,\theta}, \psi^0_{u,\theta})$ maximizes the dual objective, so does $(\varphi_{u,\theta}, \psi_{u,\theta})$. Therefore, $\varphi_{u,\theta}, \psi_{u,\theta}$ are optimal potentials. □

The following result and its proof follow Proposition 1 in Mena and Niles-Weed (2019).

Proposition 2. *For any $u \in \mathbb{B}$ and $\theta \in \Theta$, there exist optimal dual potentials $\varphi_{u,\theta}, \psi_{u,\theta}$ such that for any multi-index $\alpha \in \mathbb{N}_0^d$ of order $|\alpha| \le s$,*

$$|\nabla^\alpha \varphi_{u,\theta}(x)| \le C_{s,d} \text{ for all } x \in \mathcal{X}, \tag{10}$$

$$|\nabla^\alpha \psi_{u,\theta}(y)| \le C_{s,d} \text{ for all } y \in \mathcal{Y}, \tag{11}$$

where $C_{s,d} < \infty$ is a quantity that does not depend on x, y, u, θ, but may depend on $s, d, \bar{\phi}$, and $\bar{\phi}_s$.

Proof. Let $\varphi_{u,\theta}$ be a potential as in Proposition 1.

Denote $k = |\alpha|$. By the multivariate Faà di Bruno's formula (see, e.g., Corollary 2.10 in Constantine and Savits (1996)),

$$\nabla^\alpha \varphi_{u,\theta}(x) = \sum_{\beta_1+\cdots+\beta_k=\alpha} \lambda_{\alpha,\beta_1,\ldots,\beta_k} \prod_{j=1}^{k} \int e^{-\psi^0_{u,\theta}(y)} \nabla^{\beta_j} e^{-u'\phi(x,y,\theta)} \, d\nu(y), \tag{12}$$

where the summation is over multi-indices $\beta_1, \ldots, \beta_k \in \mathbb{N}_0^d$ such that $\beta_1 + \cdots + \beta_k = \alpha$ and $\lambda_{\alpha,\beta_1,\ldots,\beta_k}$ are combinatorial quantities that only depend on $\alpha, \beta_1, \ldots, \beta_k$.

Now consider the expression $\nabla^\beta e^{-u'\phi(x,y,\theta)}$ for a multi-index β of order $|\beta| \leq s$. Again, by Faà di Bruno's formula,

$$\nabla^\beta e^{-u'\phi(x,y,\theta)} = \sum_{\gamma_1+\cdots+\gamma_{|\beta|}=\beta} \nu_{\beta,\gamma_1,\ldots,\gamma_{|\beta|}} \prod_{j=1}^{|\beta|} \nabla^{\gamma_j} [u'\phi(x,y,\theta)]$$

for combinatorial quantities $\nu_{\beta,\gamma_1,\ldots,\gamma_{|\beta|}}$ that only depend on $\beta, \gamma_1, \ldots, \gamma_{|\beta|}$. Let $\preceq$ denote inequality up to a multiplicative constant that depends only on s and d. Then

$$\left| \nabla^\beta e^{-u'\phi(x,y,\theta)} \right| \preceq \bar{\phi}_s^{|\beta|} \preceq \max(\bar{\phi}_s, \bar{\phi}_s^s),$$

where the last inequality holds because $|\beta| \leq s$.

Combining with (12) yields

$$|\nabla^\alpha \varphi_{u,\theta}(x)| \preceq \sum_{\beta_1+\cdots+\beta_k=\alpha} |\lambda_{\alpha,\beta_1,\ldots,\beta_k}| \prod_{j=1}^{k} \bar{\phi}_s^s \int e^{-\psi^0_{u,\theta}(y)} \, d\nu(y) \preceq e^{k\bar{\phi}} \bar{\phi}_s^{ks} \preceq e^{s\bar{\phi}} \max(\bar{\phi}_s^s, \bar{\phi}_s^{s^2}),$$

where the last inequality holds because $k \leq s$. The proof for the potential $\psi_{u,\theta}$ is analogous. □

The following proposition and its proof follow Lemma E.23 of Goldfeld et al. (2024).

Proposition 3. *The mapping $\mu \mapsto c(\mu)$ is Lipschitz continuous and Gateaux differentiable, i.e., for any probability measures μ_0, μ_1 supported in $\mathcal{X}$, we have*

$$\|c(\mu_1) - c(\mu_0)\|_{\mathbb{B}\times\Theta} \leq \|\mu_1 - \mu_0\|_{\mathscr{F}} \tag{13}$$

and

$$\lim_{t\downarrow 0} \frac{c(\mu_0 + t(\mu_1 - \mu_0))(u,\theta) - c(\mu_0)(u,\theta)}{t} = \int \varphi_{u,\theta}\, d(\mu_1 - \mu_0), \tag{14}$$

where $\varphi_{u,\theta}$ is an optimal potential for μ_0 and the cost function $u'\phi(\cdot,\cdot,\theta)$.

Proof. We first prove (13). Let $\mu_t = \mu + t(\mu_1 - \mu_0)$ for $t \in [0,1]$ and let $\varphi^t_{u,\theta}$, $\psi^t_{u,\theta}$ be optimal potentials for μ_t and the cost function $u'\phi(\cdot,\cdot,\theta)$. Since μ_t is concentrated in $\mathcal{X}$, we can choose these potentials to satisfy the derivative bounds. Notice that

$$\begin{aligned} c(\mu_t)(u,\theta) &= \int \varphi^t_{u,\theta}\, d\mu_t + \int \psi^t_{u,\theta}\, d\nu \geq \int \varphi^0_{u,\theta}\, d\mu_t + \int \psi^0_{u,\theta}\, d\nu - \int e^{\varphi^0_{u,\theta}\oplus\psi^0_{u,\theta} - u'\phi(\cdot,\cdot,\theta)}\, d\mu_t \otimes \nu + 1 \\ &= \int \varphi^0_{u,\theta}\, d\mu_t + \int \psi^0_{u,\theta}\, d\nu = \int \varphi^0_{u,\theta}\, d\mu_0 + \int \psi^0_{u,\theta}\, d\nu + t\int \varphi^0_{u,\theta}\, d(\mu_1 - \mu_0) \\ &= c(\mu_0)(u,\theta) + t\int \varphi^0_{u,\theta}\, d(\mu_1 - \mu_0), \end{aligned} \tag{15}$$

where the second equality uses the fact that $\int e^{\varphi^0_{u,\theta}(x) + \psi^0_{u,\theta}(y) - u'\phi(x,y,\theta)} d\nu(y) = 1$ for all $x \in \mathcal{X}$. Hence,

$$\liminf_{t\downarrow 0} \frac{c(\mu_t)(u,\theta) - c(\mu_0)(u,\theta)}{t} \geq \int \varphi^0_{u,\theta}\, d(\mu_1 - \mu_0).$$

Similarly, we have

$$\begin{aligned} c(\mu_t)(u,\theta) &= \int \varphi^t_{u,\theta}\, d\mu_t + \int \psi^t_{u,\theta}\, d\nu = \int \varphi^t_{u,\theta}\, d\mu_0 + \int \psi^t_{u,\theta}\, d\nu + t\int \varphi^t_{u,\theta}\, d(\mu_1 - \mu_0) \\ &\leq \int \varphi^0_{u,\theta}\, d\mu_0 + \int \psi^0_{u,\theta}\, d\nu + t\int \varphi^t_{u,\theta}\, d(\mu_1 - \mu_0) + \int e^{\varphi^t_{u,\theta}\oplus\psi^t_{u,\theta} - u'\phi(\cdot,\cdot,\theta)}\, d\mu_0 \otimes \nu - 1 \\ &= \int \varphi^0_{u,\theta}\, d\mu_0 + \int \psi^0_{u,\theta}\, d\nu + t\int \varphi^t_{u,\theta}\, d(\mu_1 - \mu_0) \\ &= c(\mu_0)(u,\theta) + t\int \varphi^t_{u,\theta}\, d(\mu_1 - \mu_0), \end{aligned} \tag{16}$$

where the second equality uses the fact that $\int e^{\varphi^0_{u,\theta}(x) + \psi^0_{u,\theta}(y) - u'\phi(x,y,\theta)} d\mu_0(x) = 1$ for all $y \in \mathcal{Y}$. Hence,

$$\frac{c(\mu_t)(u,\theta) - c(\mu_0)(u,\theta)}{t} \leq \int \varphi^t_{u,\theta}\, d(\mu_1 - \mu_0).$$

It suffices to show that for any sequence $t_n \downarrow 0$,

$$\lim_{n\to\infty} \int \varphi^{t_n}_{u,\theta}\, d(\mu_1 - \mu_0) = \int \varphi^0_{u,\theta}\, d(\mu_1 - \mu_0). \tag{17}$$

Pick any subsequence n' of n. From (10) and Arzela-Ascoli theorem, there exists a further subsequence n'' such that $\varphi_{u,\theta}^{t_{n''}} \to \varphi_{u,\theta}$ and $\psi_{u,\theta}^{t_{n''}} \to \psi_{u,\theta}$ locally uniformly for some continuous functions $\varphi_{u,\theta}, \psi_{u,\theta}$. Again, from (10) and the dominated convergence theorem, $(\varphi_{u,\theta}, \psi_{u,\theta})$ satisfy the first-order conditions, and hence they are optimal potentials for (μ_0, ν) and the cost function $u'\phi(\cdot,\cdot,\theta)$. Let us now verify that $\varphi_{u,\theta} = \varphi_{u,\theta}^0 + a$ for some constant $a \in \mathbb{R}$. By uniqueness of optimal potentials, $\psi_{u,\theta} = \psi_{u,\theta}^0 - a$ for some $a \in \mathbb{R}$. Then,

$$\varphi_{u,\theta}(x) = -\log \int e^{\psi_{u,\theta}(y) - u'\phi(x,y,\theta)}\, d\nu(y) = -\log \int e^{\psi_{u,\theta}^0(y) - u'\phi(x,y,\theta)}\, d\nu(y) + a = \varphi_{u,\theta}^0(x) + a.$$

Therefore, by (10) and the dominated convergence theorem,

$$\lim_{n''\to 0} \int \varphi_{u,\theta}^{t_{n''}}\, d(\mu_1 - \mu_0) = \int \varphi_{u,\theta}\, d(\mu_1 - \mu_0) = \int \varphi_{u,\theta}^0\, d(\mu_1 - \mu_0).$$

Since the limit does not depend on the choice of the subsequence, this establishes (17), completing the proof of (14).

Next, we show (13). From the inequalities (15) and (16), we have

$$|c(\mu_1)(u,\theta) - c(\mu_0)(u,\theta)| \le \max\left(\int \varphi_{u,\theta}^0\, d(\mu_1 - \mu_0),\ \int \varphi_{u,\theta}^1\, d(\mu_1 - \mu_0) \right) \le \|\mu_1 - \mu_0\|_{\mathscr{F}},$$

where the last inequality is due to $\varphi_{u,\theta}^0, \varphi_{u,\theta}^1 \in \mathscr{F}$. Since $\mathscr{F}$ is independent of (u,θ), the proof is completed. □

Proposition 4. *There exists a tight Gaussian process $\mathbb{G}_{\mu\otimes\nu}$ in $\ell^\infty(\mathscr{F}^\oplus)$ such that*

$$\sqrt{n}(\hat{\mu}_n \otimes \hat{\nu}_n - \mu \otimes \nu) \rightsquigarrow \mathbb{G}_{\mu\otimes\nu} \text{ in } \ell^\infty(\mathscr{F}^\oplus).$$

Proof. See the proof of part (ii) of Theorem 1 in Goldfeld et al. (2024). □

A.4 Proof of Corollary 1

Theorem 3.1 in Shapiro (1991) implies that the functional $\chi(c) = \max_{u\in\mathbb{B}} c_\theta(u)$ is Hadamard directionally differentiable with the derivative

$$\chi_c'(h) = \max_{u\in U_c} h(u), \quad h \in C(\mathbb{B}),$$

where

$$U_c = \arg\max_{u\in\mathbb{B}} c_\theta(u).$$

Applying the functional delta method (e.g., van der Vaart and Wellner, 2023, Theorem 3.10.5) to Theorem 3 completes the proof.

B Details for panel logit with attrition and refreshment

B.1 Common slope parameter

For the fixed effects panel logit with $T = 2$, the conditional log-likelihood for individuals with $S_i = Y_{i1} + Y_{i2} = 1$ is

$$\ell_i(\theta \mid S_i = 1) = Y_{i1}X_{i1}'\theta + Y_{i2}X_{i2}'\theta - \ln\left(e^{X_{i1}'\theta} + e^{X_{i2}'\theta}\right),$$

with score

$$s(Y_{i1}, Y_{i2}, X_{i1}, X_{i2}; \theta) = (Y_{i1}X_{i1} + Y_{i2}X_{i2}) - \frac{e^{X_{i1}'\theta}X_{i1} + e^{X_{i2}'\theta}X_{i2}}{e^{X_{i1}'\theta} + e^{X_{i2}'\theta}},$$

and moment condition

$$\mathbb{E}\left[s(Y_{i1}, Y_{i2}, X_{i1}, X_{i2}; \theta_0) \mid S_i = 1\right] = 0.$$

We embed the event $\{Y_{i1} + Y_{i2} = 1\}$ in the cost function

$$\phi\left(y_1, y_2, x_1, x_2; \theta\right) = s(y_1, y_2, x_1, x_2; \theta)\mathbf{1}\left\{y_1 + y_2 = 1\right\}.$$

We partition the population into retainers (observed in both periods) and attriters (observed only in period 1). This partitioned approach yields tighter bounds by fixing the known joint distribution of retainers and limiting the OT problem to the attriters only. Let p denote the retention rate. We use the following notations.

- $f_1(y, x)$: distribution of all units in period 1,
- $f_{1|\text{ret}}(y, x)$: distributions of retainers in period 1,
- $f_{1|\text{att}}(y, x)$: distributions of attriters in period 1,
- $f_{2|\text{ref}}(y, x) = f_2(y, x)$: distribution of the refreshment sample in period 2, which equals the unconditional distribution in period 2 since the refreshment sample is drawn from the same population as the original sample,
- $f_{2|\text{ret}}(y, x)$: distribution of retainers in period 2,
- $f_{1,2|\text{ret}}(y_1, y_2, x_1, x_2)$: joint distribution of retainers for both periods,

- $f_{2|\text{att}}(y, x)$: distribution of attriters in period 2 (unobserved).

By the law of total probability,

$$f_1 = p \cdot f_{1|\text{ret}} + (1-p) \cdot f_{1|\text{att}},$$
$$f_2 = p \cdot f_{2|\text{ret}} + (1-p) \cdot f_{2|\text{att}},$$

so the unobserved attriter distribution in period 2 is

$$f_{2|\text{att}} = \frac{f_2 - p \cdot f_{2|\text{ret}}}{1-p}.$$

Our OT problem couples only the attriter distributions $f_{1|\text{att}}$ and $f_{2|\text{att}}$, while the joint distribution of retainers $f_{1,2|\text{ret}}$ is fixed at its observed value. Let $\Pi(f_{1|\text{att}}, f_{2|\text{att}})$ denote the set of all joint distributions with these marginals. The attriter contribution to the moment bounds is

$$\underline{\nu}_{\text{att}}(\theta) = \inf_{f \in \Pi(f_{1|\text{att}}, f_{2|\text{att}})} \mathbb{E}_f[\phi(Y_1, Y_2, X_1, X_2; \theta)],$$
$$\overline{\nu}_{\text{att}}(\theta) = \sup_{f \in \Pi(f_{1|\text{att}}, f_{2|\text{att}})} \mathbb{E}_f[\phi(Y_1, Y_2, X_1, X_2; \theta)],$$

and the overall bounds are obtained by combining the contributions from retainers and attriters

$$\underline{\nu}(\theta) = p \cdot \mathbb{E}_{f_{1,2|\text{ret}}}[\phi(Y_1, Y_2, X_1, X_2; \theta)] + (1-p) \cdot \underline{\nu}_{\text{att}}(\theta),$$
$$\overline{\nu}(\theta) = p \cdot \mathbb{E}_{f_{1,2|\text{ret}}}[\phi(Y_1, Y_2, X_1, X_2; \theta)] + (1-p) \cdot \overline{\nu}_{\text{att}}(\theta).$$

Note that given the indicator function $\mathbf{1}\{Y_1 + Y_2 = 1\}$ in the cost function $\phi(y_1, y_2, x_1, x_2; \theta)$, the OT coupling allocates maximal mass to zero-cost non-switchers at each covariate value x,

$$w_{00}(x) = \min\left\{f_{1|\text{att}}(0, x), f_{2|\text{att}}(0, x)\right\},$$
$$w_{11}(x) = \min\left\{f_{1|\text{att}}(1, x), f_{2|\text{att}}(1, x)\right\}.$$

The remaining mass on the informative switcher pairs $(y_1, y_2) = (0, 1)$ or $(1, 0)$ is

$$1 - w_{00}(x) - w_{11}(x) = \left|f_{1|\text{att}}(1, x) - f_{2|\text{att}}(1, x)\right|.$$

If the marginal distributions are identical across periods, this quantity is zero, meaning no switchers exist and θ cannot be identified under unrestricted attrition.

Algorithm 4 presents the algorithm for computing bounds on the common slope parameter θ. In line 4, we estimate $\hat{f}_{2|\text{ret}}$ using a nonparametric kernel estimator since it must be evaluated at the

refreshment sample points during the OT computation in lines 10–11, where $k(\cdot)$ is a kernel function and h is the bandwidth. Alternative nonparametric estimators, such as sieves or splines, could be employed as well. Intuitively, the OT coupling in lines 10–11 effectively reweights the refreshment sample $\{(Y_{j2}, X_{j2})\}$ to approximate the unobserved attriter distribution $\hat{f}_{2|\text{att}}$.

B.2 AME

B.2.1 Without attrition

Davezies et al. (2024) use Chebyshev polynomial approximation to construct outer bounds on the AME.[1] Let $X = (X_1, \ldots, X_T)$ denote the full stack of period-specific covariates and $S_i = \sum_{t=1}^{T} Y_{it}$. The AME of covariate j at period τ is

$$\delta_{\tau,j} = \theta_j \, \mathbb{E}[\Lambda(X_\tau'\theta + \alpha)(1 - \Lambda(X_\tau'\theta + \alpha))].$$

The outer bounds of $\delta_{\tau,j}$ are given by $\tilde{\delta} \pm \bar{b}$, where $\tilde{\delta} = \mathbb{E}[p(X, S, \theta_0)]$ and $\bar{b} = \mathbb{E}[a(X, S, \theta_0)]$, with

$$\begin{aligned}
p(x, s, \theta) &= \sum_{t=0}^{s} \left(\lambda_t(x, \theta) + b^*_{t,T}\lambda_{T+1}(x, \theta)\right) \binom{T-t}{s-t} \frac{\exp(sx_\tau'\theta)}{C_s(x, \theta)}, \\
a(x, s, \theta) &= \frac{1}{2 \times 4^T} |\lambda_{T+1}(x, \theta)| \binom{T}{s} \frac{\exp(sx_\tau'\theta)}{C_s(x, \theta)}, \\
C_k(x, \theta) &= \sum_{(d_1,\ldots,d_T)\in\{0,1\}^T : \sum_{t=1}^{T} d_t = k} \exp\left(\sum_{t=1}^{T} d_t x_t'\theta\right).
\end{aligned}$$

For notational simplicity, we suppress the subscripts τ and j when there is no ambiguity. Here $b^*_{t,T}$ are the Chebyshev coefficients that are fixed constants fully determined by the degree-$(T+1)$ Chebyshev polynomial rescaled to $[0, 1]$, independent of any data. The functions $\lambda_t(x, \theta)$ are defined as the monomial coefficients in

$$\sum_{t=0}^{T+1} \lambda_t(x, \theta) u^t = \theta_j u(1-u) \prod_{t\neq\tau} \left(1 + u\left\{\exp\left((x_t - x_\tau)'\theta\right) - 1\right\}\right).$$

For simplicity, we now illustrate the case with $T = 2$. Then

$$\begin{aligned}
C_0(x, \theta) &= \exp(0) = 1, \\
C_1(x, \theta) &= \exp(x_1'\theta) + \exp(x_2'\theta), \\
C_2(x, \theta) &= \exp(x_1'\theta + x_2'\theta).
\end{aligned}$$

[1] A sharper bound can be obtained via Hankel moment matrix positivity, but it involves nonparametric first-step estimation.

Algorithm 4 Panel logit with attrition and refreshment: common slope parameter

Require: Original panel at $t = 1$: $\{(Y_{i1}, X_{i1})\}_{i=1}^{n_{\text{org}}}$; Retainers' sample at $t = 2$: $\{(Y_{i2}, X_{i2})\}_{i \in \mathcal{S}}$; Refreshment sample at $t = 2$: $\{(Y_{j2}, X_{j2})\}_{j=1}^{n_{\text{ref}}}$

Ensure: Identified set $\widehat{\Theta}_I$ for parameter θ

$n_{\text{ret}} \leftarrow |\mathcal{S}|$, $\hat{p} \leftarrow n_{\text{ret}}/n_{\text{org}}$

Estimate empirical marginals:

$\hat{f}_{1|\text{ret}}(y, x) \leftarrow \frac{1}{n_{\text{ret}}} \sum_{i \in \mathcal{S}} \mathbf{1}\{Y_{i1} = y, X_{i1} = x\}$

$\hat{f}_{1|\text{att}}(y, x) \leftarrow \frac{1}{n_{\text{org}} - n_{\text{ret}}} \sum_{i \notin \mathcal{S}} \mathbf{1}\{Y_{i1} = y, X_{i1} = x\}$

$\hat{f}_{2|\text{ret}}(y, x) \leftarrow \frac{1}{n_{\text{ret}} h^d} \sum_{i \in \mathcal{S}} \mathbf{1}\{Y_{i2} = y\} k\left(\frac{X_{i2} - x}{h}\right)$

$\hat{f}_2(y, x) \leftarrow \frac{1}{n_{\text{ref}}} \sum_{j=1}^{n_{\text{ref}}} \mathbf{1}\{Y_{j2} = y, X_{j2} = x\}$

Recover attriter marginal:

$\hat{f}_{2|\text{att}}(y, x) \leftarrow \frac{\hat{f}_2(y,x) - \hat{p}\hat{f}_{2|\text{ret}}(y,x)}{1 - \hat{p}}$

Construct cost matrix:

for each $i \notin \mathcal{S}$, $j = 1, \ldots, n_{\text{ref}}$ **do**

  $\phi_{ij}(\theta) \leftarrow s(Y_{i1}, Y_{j2}, X_{i1}, X_{j2}; \theta) \mathbf{1}\{Y_{i1} + Y_{j2} = 1\}$

end for

Solve entropic OT problems:

$\widehat{\underline{\nu}}_{\text{att}}(\theta) \leftarrow \min_{w \geq 0} \sum_{i,j} w_{ij} \phi_{ij}(\theta) + \varepsilon \sum_{i,j} w_{ij} \log \frac{w_{ij}}{\hat{f}_{1|\text{att}}(Y_{i1}, X_{i1}) \hat{f}_{2|\text{att}}(Y_{j2}, X_{j2})}$ subject to constraints

$\widehat{\overline{\nu}}_{\text{att}}(\theta) \leftarrow \max_{w \geq 0} \sum_{i,j} w_{ij} \phi_{ij}(\theta) + \varepsilon \sum_{i,j} w_{ij} \log \frac{w_{ij}}{\hat{f}_{1|\text{att}}(Y_{i1}, X_{i1}) \hat{f}_{2|\text{att}}(Y_{j2}, X_{j2})}$ subject to constraints

where constraints are $\sum_j w_{ij} = \hat{f}_{1|\text{att}}(Y_{i1}, X_{i1})$, $\sum_i w_{ij} = \hat{f}_{2|\text{att}}(Y_{j2}, X_{j2})$

Compute retainer moment:

$\widehat{\nu}_{\text{ret}}(\theta) \leftarrow \frac{1}{n_{\text{ret}}} \sum_{i \in \mathcal{S}} \phi(Y_{i1}, Y_{i2}, X_{i1}, X_{i2}; \theta)$

Combine bounds:

$\widehat{\underline{\nu}}(\theta) \leftarrow \hat{p}\widehat{\nu}_{\text{ret}}(\theta) + (1 - \hat{p})\widehat{\underline{\nu}}_{\text{att}}(\theta)$

$\widehat{\overline{\nu}}(\theta) \leftarrow \hat{p}\widehat{\nu}_{\text{ret}}(\theta) + (1 - \hat{p})\widehat{\overline{\nu}}_{\text{att}}(\theta)$

Construct identified set:

$\widehat{\Theta}_I \leftarrow \{\theta : \widehat{\underline{\nu}}(\theta) \leq 0 \leq \widehat{\overline{\nu}}(\theta)\}$

return $\widehat{\Theta}_I$

For $\lambda_t(x,\theta)$, w.l.o.g. let $\tau = 1$. Due to the factor $u(1-u)$, we have $\lambda_0(x,\theta) = 0$. Then, expanding the defining identity gives

$$\lambda_1(x,\theta)u + \lambda_2(x,\theta)u^2 + \lambda_3(x,\theta)u^3 = \theta_j u(1-u)\left(1 + u\left\{\exp\left((x_2 - x_1)'\theta\right) - 1\right\}\right),$$

which implies

$$\lambda_1(x,\theta) = \theta_j, \quad \lambda_2(x,\theta) = \theta_j \exp\left((x_2 - x_1)'\theta\right) - 2, \quad \lambda_3(x,\theta) = \theta_j\left(1 - \exp\left((x_2 - x_1)'\theta\right)\right).$$

Using these explicit forms, we can simplify the expressions for $p(x,s,\theta)$ and $a(x,s,\theta)$. For $S = 0$,

$$p(x,0,\theta) = b^*_{0,2}\lambda_3(x,\theta), \quad a(x,0,\theta) = \frac{1}{32}|\lambda_3(x,\theta)|.$$

For $S = 1$,

$$\begin{aligned} p(x,1,\theta) &= \left[\theta_j + (2b^*_{0,2} + b^*_{1,2})\lambda_3(x,\theta)\right]\frac{\exp(x'_\tau\theta)}{\exp(x'_1\theta) + \exp(x'_2\theta)}, \\ a(x,1,\theta) &= \frac{1}{16}|\lambda_3(x,\theta)|\frac{\exp(x'_\tau\theta)}{\exp(x'_1\theta) + \exp(x'_2\theta)}. \end{aligned}$$

For $S = 2$,

$$\begin{aligned} p(x,2,\theta) &= (b^*_{0,2} + b^*_{1,2} + b^*_{2,2} - 1)\lambda_3(x,\theta)\frac{\exp(2x'_\tau\theta)}{\exp(x'_1\theta + x'_2\theta)}, \\ a(x,2,\theta) &= \frac{1}{32}|\lambda_3(x,\theta)|\frac{\exp(2x'_\tau\theta)}{\exp(x'_1\theta + x'_2\theta)}. \end{aligned}$$

By plugging in $\hat{\theta}$ and forming the corresponding sample analogues, one obtains the estimates $\tilde{\delta}$ and $\bar{b}$ used in the AME bounds in the main text. Davezies et al. (2024) also show how to construct valid confidence intervals based on these bounds.

B.2.2 With unrestricted attrition

To obtain AME bounds under unrestricted attrition, we proceed in three steps.

Step 1: bounds for θ. Compute the identified set $\widehat{\Theta}_I$ for the common slope parameters θ as described above and in Algorithm 4. Let $\left\{\theta^{(g)}\right\}_{g=1}^{G}$ be a finite grid covering $\widehat{\Theta}_I$.

Step 2: bounds for the AME conditional on $\theta^{(g)}$. For each grid point, plug in $\theta^{(g)}$ into the Chebyshev polynomial approximation and define cost functions

$$\left[\underline{\phi}(x,s,\theta^{(g)}),\, \overline{\phi}(x,s,\theta^{(g)})\right] = p(x,s,\theta^{(g)}) \pm a(x,s,\theta^{(g)}),$$

where p and a are as in Appendix B. We then solve OT problems for the attriter sample,

$$\underline{\delta}_{\text{att}}(\theta^{(g)}) = \inf_{f \in \Pi(f_{1|\text{att}}, f_{2|\text{att}})} \mathbb{E}_f \left[\underline{\phi}(X, S, \theta^{(g)}) \right],$$
$$\overline{\delta}_{\text{att}}(\theta^{(g)}) = \sup_{f \in \Pi(f_{1|\text{att}}, f_{2|\text{att}})} \mathbb{E}_f \left[\overline{\phi}(X, S, \theta^{(g)}) \right].$$

The identified set for the AME under $\theta^{(g)}$ is $\left[\underline{\delta}(\theta^{(g)}), \overline{\delta}(\theta^{(g)})\right]$, where we combine retainers and attriters as

$$\underline{\delta}(\theta^{(g)}) = p \cdot \mathbb{E}_{\text{ret}} \left[\underline{\phi}(X, S, \theta^{(g)}) \right] + (1-p) \cdot \underline{\delta}_{\text{att}}(\theta^{(g)}),$$
$$\overline{\delta}(\theta^{(g)}) = p \cdot \mathbb{E}_{\text{ret}} \left[\overline{\phi}(X, S, \theta^{(g)}) \right] + (1-p) \cdot \overline{\delta}_{\text{att}}(\theta^{(g)}),$$

and $\mathbb{E}_{\text{ret}} \left[\underline{\phi}(X, S, \theta^{(g)}) \right]$ and $\mathbb{E}_{\text{ret}} \left[\overline{\phi}(X, S, \theta^{(g)}) \right]$ can be computed directly from the observed data for retainers.

Step 3: profiling over θ. Finally, we take the union over the grid to obtain the identified set for the AME

$$\bigcup_{g=1}^{G} \left[\underline{\delta}(\theta^{(g)}), \overline{\delta}(\theta^{(g)}) \right].$$