Asymptotic Minimax Optimality of Set Estimation

Hirotaka Ito*
Universitat Pompeu Fabra

Job Market Paper
Click here for the latest version.

First version: October 4, 2026, This version: November 13, 2026

Abstract

Non-convex identified sets of parameters arise naturally in econometric models through nonlinear relationships and structural identifying restrictions. Yet existing efficiency theory for set estimation relies on convexity, leaving open the fundamental question of how accurately these sets can be estimated from data. This paper develops a general theory of asymptotically efficient set estimation that accommodates a wider class of identified sets. I first derive the local asymptotic minimax efficiency bound under risk functions based on the Hausdorff metric. A key ingredient in the analysis is a limit experiment framework tailored to the local geometry of the identified set. I then establish the limiting distribution of a set estimator based on a criterion function. Unlike existing approaches that rely on support functions, the analysis preserves the full geometry of non-convex sets. The limiting distribution motivates a bootstrap-based confidence set, whose finite-sample performance is examined through Monte Carlo simulations. An application to optimal top income taxation with behavioral agents illustrates the empirical relevance of my framework.

Keywords: partial identification, set estimation, asymptotic optimality.
JEL Classification Codes: C10, C13, C14, C18.

*Department of Economics and Business, Universitat Pompeu Fabra, Spain. Address: Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain. Email: hirotaka.ito@upf.edu.

I am deeply grateful to Geert Mesters for his support and guidance. I sincerely thank Kirill S. Evdokimov, Toru Kitagawa and Adam Lee for their helpful comments and discussions. All errors are mine.

1. Introduction

When a parameter of interest is potentially set identified, what is the best possible precision with which a researcher can estimate such a set? Since Manski (1989), this question has been central to empirical research that relies on partial identification. Here the identified set is the set of parameters that are consistent with the assumptions or restrictions imposed on the data and the econometric model. Accordingly, the geometry of the identified set is economically informative; its size captures the identification strength of the assumptions employed, while its shape reflects the geometric nature of the restrictions (e.g., disconnected regions, curved boundaries, and non-convexities), revealing which combinations of economically relevant parameters are admissible. Thus, it is important to estimate the identified set as precisely as possible while preserving its geometric structure. This is a question of efficiency for a set-identified object.

Regarding the statistical noise for set-valued objects, little is known about how well we can handle it. This is in sharp contrast to the point-identified case. When a parameter is point identified, classical theory provides a lower bound on such noise; examples include the Cramér–Rao bound, the Hájek–Le Cam convolution and local asymptotic minimax theorems (e.g., Hájek, 1970; Hájek, 1972; Le Cam, 1972), and their counterparts for moment condition models (e.g., Hansen, 1982; Chamberlain, 1987). When a parameter is set identified, the corresponding theory is much more limited. Kaido and Santos (2014) establish an efficiency bound for convex identified sets in a class of moment inequality models. Their approach exploits the fact that a convex set can be fully characterized by its support function, which records the maximal value of linear projections of the set in different directions. This representation allows the estimation of a convex set to be studied through the estimation of its support function. However, this approach is not suitable when the identified set is non-convex, because the estimator necessarily recovers the convex hull of that set, completely breaking the original shape, which incorporates the information in the econometric restrictions.

This paper develops an asymptotic efficiency theory for the estimation of general identified sets. I make two main methodological contributions. First, I derive a local asymptotic minimax lower bound for set estimation under loss functions based on the Hausdorff metric. The bound builds on a limit experiment that captures how the statistical information for estimating a set varies across its boundary and across directions of local perturbation. The framework applies to a broad class of compact identified sets, including finite collections of points, disconnected regions, and sets with piecewise smooth boundaries and corners; since a singleton is naturally included in my framework, classical point-estimation efficiency is derived as a special case. Second, I derive the limiting distribution, in the Hausdorff metric, of a set estimator based on a criterion function. Unlike existing approaches that characterize the target set through its support function, my analysis directly tracks deviations between the estimator and the population set. This local analysis bypasses the global characterization of the identified set and allows me to preserve non-convexity.

Empirical relevance. Non-convex identified sets arise naturally in a wide range of empirical problems. Two broad classes can be distinguished.

First, economic models frequently impose nonlinear relationships between parameters, and these restrictions can generate non-convex sets of admissible parameters even when the underlying primitives belong to convex sets. Examples include optimal top income taxation, where the optimal rate depends nonlinearly on the income distribution and on behavioral responses (Saez, 2001; Farhi and Gabaix, 2020); production-based markups, which depend on output elasticities and input revenue shares (De Loecker and Warzynski, 2012; De Loecker et al., 2020; Bond et al., 2021); and optimal unemployment insurance, which balances consumption-smoothing benefits against behavioral responses (Gruber, 1997; Chetty, 2008). Similar nonlinear relationships arise in the marginal value of public funds (Hendren and Sprung-Keyser, 2020), the value of a statistical life (Ashenfelter and Greenstone, 2004), implied discount rates in energy-efficiency decisions (Busse et al., 2013; Allcott and Wozny, 2014), sensitivity analysis to unobserved selection (Oster, 2019), and turning points in inverted-U relationships (Aghion et al., 2005). When the underlying parameters are partially identified, these relationships can generate non-convex joint identified sets for the economically relevant objects.

Second, non-convexity also arises from the identifying restrictions imposed by structural economic models. In dynamic stochastic general equilibrium (DSGE) models, nonlinear equilibrium conditions and cross-equation restrictions link structural parameters to observable economic dynamics. When these parameters are not uniquely identified, the resulting identified sets can be non-convex or disconnected (Canova and Sala, 2009; Komunjer and Ng, 2011). Similarly, in structural VARs, orthogonality and sign restrictions on structural shocks can generate non-convex joint identified sets for impulse responses across variables and horizons (Uhlig, 2005; Moon and Schorfheide, 2012; Baumeister and Hamilton, 2015; Granziera et al., 2018; Giacomini and Kitagawa, 2021).

These examples demonstrate that non-convexity is not a technical curiosity but a consequence of economically meaningful restrictions. Replacing the identified set by its convex hull can introduce parameter combinations that violate the underlying econometric restrictions, potentially distorting economic conclusions. Moreover, since the Hausdorff metric is not invariant under nonlinear transformations, efficient estimation of partially identified primitives does not generally imply efficient estimation of the target set. A general efficiency theory must therefore accommodate the full geometry of the identified set.

Theoretical framework. My theoretical framework proceeds as follows. Let $\Theta_I \subset \mathbb{R}^{d_\theta}$ denote the identified set, and let T_n be a set-valued estimator based on a sample of size n. I measure estimation error using the Hausdorff distance $d_H(T_n, \Theta_I)$, which controls both the largest distance from an estimated point to the true set and the largest distance from a true point to the estimated set. This two-sided metric is particularly appropriate for non-convex targets: it detects missing components, spurious regions, and incorrect boundary shapes that a support-function representation can conceal. The object of the efficiency analysis is the local asymptotic minimax risk under loss functions of the form $l\big(\sqrt{n} \cdot d_H(T_n, \Theta_I)\big)$, where l is a suitable function.

The first theoretical contribution constructs a limit experiment for partially identified models. I extend the limit-experiment framework of Le Cam (1972) to the case in which multiple parameter values are observationally equivalent. The crucial feature is that local information is not summarized by a single score or information matrix. At a point on the boundary $\partial\Theta_I$, the effect of a local perturbation can depend on its direction: some directions can remain within the observationally equivalent region, while others move outside it and are statistically distinguishable. I capture this feature through direction-dependent scores and local information. The resulting limit experiment is represented by a Gaussian field indexed by boundary points and local directions.

I use this limit experiment to derive a lower bound on the asymptotic risk of set-valued estimators. The bound is determined by the local information available along the boundary and by the directions in which that information is weakest. This approach extends the local asymptotic minimax logic of classical point estimation to the set-valued problem without requiring global convexity. I first establish the result for a parametric model and then extend it to general moment-restriction models using the semiparametric approach of Chamberlain (1987). When the identified set is a singleton, the framework nests the familiar efficiency theory for point-identified parameters.

The second theoretical contribution characterizes the distribution of a feasible estimator. I consider sets defined by moment inequalities and estimate them as lower contour sets of a sample criterion function, following the criterion-function approach of Manski and Tamer (2002) and Chernozhukov et al. (2007). Under regularity conditions on the moment restrictions and their active gradients, I derive the limiting distribution of $\sqrt{n} \cdot d_H(\widehat{\Theta}_I, \Theta_I)$. The distribution is a functional of a Gaussian field evaluated on $\partial\Theta_I$, with the binding restrictions determining how sampling fluctuations shift each part of the boundary. The result directly accommodates curved boundaries and corners, including regions where several inequalities bind simultaneously.

A central technical innovation is to obtain this distribution without representing Θ_I by its support function. I instead analyze the stochastic field associated with $d_H(\widehat{\Theta}_I, \Theta_I)$. Because this field is not asymptotically equicontinuous, standard weak-convergence arguments do not apply directly. I overcome this difficulty using weak epiconvergence-type arguments to track the behavior of the stochastic field, with basic techniques similar to those in Chernozhukov et al. (2007) and Kaido (2016). The resulting representation connects the sampling error of the estimated set to local boundary information while retaining the full, potentially non-convex, geometry of the target.

The limiting distribution also motivates an inference procedure. I construct a Hausdorff-ball confidence set by enlarging the estimated identified set by a radius calibrated via a multiplier bootstrap for the directed Hausdorff distance. Unlike confidence regions based on the support-function approach, this construction does not replace the target by its convex hull. I investigate its finite-sample performance in simulations. A general proof of bootstrap validity is left for future research.

Simulation evidence. I evaluate the finite-sample performance of the proposed estimator and Hausdorff-ball confidence set based on the behavioral optimal-tax model of Farhi and Gabaix (2020). The exercise is designed around the two features that motivate

the theory: partial identification arising from top-coded income data, and a non-convex identified set generated by the nonlinear relationship between the optimal tax rate and the shape of tax misperceptions. Four designs vary the attention parameter and the taxable-income elasticity, producing identified sets with different degrees of curvature. I consider samples of $n \in \{500,\ 1{,}000,\ 5{,}000\}$ and compare the proposed procedures with support-function estimation and confidence sets, subsampling methods (Chernozhukov et al., 2007), bootstrap moment-inequality inference (Bugni, 2010), and Monte Carlo confidence sets (Chen et al., 2018). The simulations deliver three main findings. First, preserving the non-convex identified set substantially improves set estimation in the Hausdorff metric relative to estimating its convex hull. This result demonstrates that convexification is not an innocuous approximation when the parameter of interest is the full identified set. Second, the asymptotic distribution derived in the paper closely approximates the finite-sample distribution of the scaled Hausdorff error, already at $n = 1{,}000$ across the four designs. Third, the bootstrap Hausdorff-ball confidence set delivers coverage close to its nominal level. Based on coverage and (median) Hausdorff distance, the finite-sample performance of the Hausdorff-ball confidence set is as good as that of its bootstrap- and Monte Carlo-based counterparts.

The evidence reinforces the paper's central message: the shape of an identified set is part of the economic object, and statistical procedures should be evaluated by how accurately they recover that shape. By developing an efficiency benchmark and a limiting distribution directly in the Hausdorff metric, the paper provides a framework for doing so without imposing convexity.

1.1. **Literature.** This paper is related to several strands of the literature. The first is the large literature on estimation and inference in partially identified models, which is now mature; see Molinari (2020) for details. The most related paper in this strand is Chernozhukov et al. (2007), who develop a general criterion function approach to set estimation and confidence regions for identified sets, establishing consistency and rates in the Hausdorff metric; the estimator I study belongs to their class. The intuition behind this study stems from the literature on (unconditional) moment inequalities (e.g., Romano and Shaikh (2010), Andrews and Guggenberger (2009), Andrews and Soares (2010), Bugni (2010), and Canay (2010)).

This paper also contributes to the literature on various notions of optimality in partially identified models. Most of this work concerns hypothesis testing, where optimality is expressed in terms of power. Andrews and Soares (2010) compare the asymptotic power of competing tests of moment inequalities; Armstrong (2014) studies power against local alternatives for conditional moment inequalities; Bugni et al. (2012) analyze how local misspecification distorts the size of these procedures. The paper closest to mine in this strand is Kaido and Santos (2014). The authors consider identified sets defined by moment inequalities that are convex in the parameter, with the slopes of any linear constraints known, so that the identified set is convex and characterized by its support function; this characterization via the support function is used in Beresteanu and Molinari (2008), Beresteanu et al. (2011), and Bontemps et al. (2012) in the econometrics literature. They apply the techniques of semiparametric efficient

estimation (e.g., Bickel et al. (1998)) to support function estimation and derive the efficiency bound. My approach is different in that I do not embed the identified set, via its support function, into the space of continuous functions indexed by the unit vectors; this global characterization of the identified set cannot distinguish two sets whose convex hulls coincide. Instead, I focus on local deviations at each point on the boundary and connect them to their implications for the Hausdorff distance. This local nature widens the scope of the efficiency bound to include finite sets of points, disconnected sets, and sets with outward kinks as special cases.

The basis of this paper's technical development is the limit experiment and statistical decision frameworks (e.g., Le Cam (1972), Hájek (1970), and van der Vaart (1991)). Both have been widely used in econometrics to evaluate estimators, tests and decision rules (e.g., Hirano and Porter (2009), Müller (2011), and Andrews and Mikusheva (2022)); see Hirano and Porter (2020) for a detailed account of the framework and its recent developments. In this literature, the object of interest is a point in a finite- or infinite-dimensional space, and efficiency results hinge on the derivative of some criterion function at that point. This approach does not extend directly to the set-identified context. I tailor the underlying limit experiment framework to partially identified models so that set estimation can be analyzed in a manner similar to the classical point-identified case.

1.2. **Organization.** The rest of the paper is organized as follows. Section 2 presents a motivating application of my framework to optimal taxation with behavioral agents. Section 3 presents the general setting and assumptions, including the class of identified sets covered in the analysis. Section 4 develops the tailored limit experiment framework and derives the lower bound for set estimators. I also extend the analysis to the case where the identified set is defined via general moment restrictions. Section 5 studies the asymptotic properties of the set estimator based on a criterion function and derives its limiting distribution in the Hausdorff metric. Section 6 presents Monte Carlo simulations. Section 7 concludes. All proofs are collected in the Appendix.

2. Motivating example

I here present the optimal top tax rate with behavioral agents as an example in which this paper can be useful for economic questions studied in the empirical literature. The logic behind this example is general enough to apply to other contexts, such as computing markups or deriving optimal unemployment benefits. The unifying theme in these examples is as follows. Many parameters of interest in applied work, such as optimal tax rates and markups, are nonlinear functions of a few primitive quantities. Two points arise. First, point identification of the primitives often relies on assumptions that are not motivated by economic theory, and dropping these assumptions leaves the primitives partially identified. Second, a nonlinear map can turn a convex set of primitives into a non-convex set of the parameter of interest. Existing optimality results for set estimation do not cover this case. Nor can one estimate the set of primitives and

then map it: the Hausdorff metric is not preserved under nonlinear transformations, so optimality of the first step does not carry over to the second.

Policy parameter of interest. I start from formally defining parameter of interest. Consider taxpayers whose income is above a threshold z^* and who face a constant marginal tax rate τ on income above z^*. A small increase in τ raises revenue from income already reported in the bracket and lowers it through the behavioral response of reported income. Saez (2001) shows that the revenue-maximizing rate is

$$\tau^* = \frac{1}{1+\pi\cdot e}, \quad \pi := \frac{z_m(z^*)}{z_m(z^*)-z^*}, \quad z_m(z^*) := \mathbb{E}_P[Z \mid Z \geq z^*], \tag{1}$$

where Z denotes reported income, π is the Pareto parameter of the upper tail, and e is the elasticity of reported income with respect to the net-of-tax rate $1-\tau$[1]. The formula (1) is a sufficient statistic result in the sense of Chetty (2009).

The formula in Saez (2001) presumes that taxpayers respond to their true marginal rate. Motivated by the large literature in behavioral economics[2], Farhi and Gabaix (2020) extend Saez (2001) and derive the optimal top rate when taxpayers misperceive the tax schedule. In their framework, a taxpayer with income z perceives the marginal rate

$$T'_{,s}(z) = \omega\cdot T'(z) + (1-\omega)\left[\int_0^\infty T'(uz)\,\psi(u)\,du + b(z)T(0)\right],$$

with $\int_{\mathbb{R}_+}\psi(u)\,du = 1$ and $\lim_{z\to\infty} b(z) = 0$, that is, a weighted average of the true marginal rate, with attention weight $\omega\in[0,1]$, and of a weighted average of the marginal rates at other income levels uz, plus a term in $T(0)$ that vanishes for large z. This represents the situation where a taxpayer could respond to a rate that is far from the true marginal rate. In the revenue-maximizing case without income effects, Proposition 12 in Farhi and Gabaix (2020) gives

$$\bar\tau = \frac{1}{1+\bar\zeta_{c,r}\cdot\pi\cdot\kappa(A)}, \quad \kappa(A) := \omega + (1-\omega)A, \quad A := \int_0^\infty u^{\pi-1}\psi(u)\,du, \tag{2}$$

where $\bar\zeta_{c,r}$ is the compensated elasticity of a taxpayer who perceives the schedule correctly. The factor κ is the fraction of a top-rate reform to which taxpayers respond, and A measures the shape of misperception. If taxpayers iron, the misperceived rate is the average rate $T(z)/z = \int_0^1 T'(uz)\,du$, so that ψ is uniform on $[0,1]$ and $A = 1/\pi < 1$. A misperceived rate that puts weight on incomes above the taxpayer's own, such as a salient statutory top rate, gives $A > 1$. Note that either $\omega = 1$ or $A = 1$ yields $\kappa = 1$; hence (2) reduces to (1) with $e = \bar\zeta_{c,r}$.

[1] Saez (2001) denotes the Pareto parameter by a. I use π to be consistent with the notation of Farhi and Gabaix (2020).

[2] De Bartolome (1995) finds in the laboratory that many subjects use the average rate as if it were the marginal rate. Feldman et al. (2016) find that households reduce reported wage income when the Child Tax Credit expires, a lump-sum change that leaves the marginal rate unchanged. Rees-Jones and Taubinsky (2020) find in a survey experiment that taxpayers commonly "iron", that is, act on their average rate, while few "spotlight" a single bracket. Attention to taxes also varies across individuals (Taubinsky and Rees-Jones, 2018).

Given the setting above, an interesting question would be how the deviation of the perceived rate from the true one affects the optimal tax rate. Since A captures the shape of the misperception, in empirical work this question boils down to finding the pairs $(\bar{\tau}, A)$ that are consistent with the available data. The object of interest is then formally expressed as the pair that is supported by the data, i.e.,

$$\theta = (\bar{\tau}, A)^t \in \Theta \subset \mathbb{R}^{d_\theta}, \quad d_\theta = 2,$$

which records which form of misperception is required for a given rate to be optimal[3]. If all ingredients in equation (2) can be uniquely pinned down from the data, then the pair $(\bar{\tau}, A)$ of interest is uniquely determined, and the identified set is a singleton in $\mathbb{R}^{d_\theta}$. However, the assumptions required for point identification of these ingredients are stringent in practice as discussed below. In this case, instead of assuming a unique pair of $(\bar{\tau}, A)$, focusing on the "profile" of the set of pairs $(\bar{\tau}, A)$ that is supported by the data and by weaker assumptions provides robust policy implications.

Partial identification. As illustration, I here focus on two inputs of (2) that are not point identified: the tail of the income distribution and the misperception shape.

The first is the tail. The mean income $z_m(z^*)$ is an average over an unbounded region, which is determined by the observations that data sources withhold; published tax tabulations have an open top bracket, and public-use survey micro-data are top-coded. Piketty and Saez (2003) address this by Pareto interpolation, which assumes the functional form whose parameter is at issue. There is no consensus regarding this point. Auten and Splinter (2024) find that top income shares are lower, and have risen less since 1980, than those in Piketty et al. (2018), and the difference turns on the allocation of income missing from tax returns.

The second is the misperception shape. While Rees-Jones and Taubinsky (2020) provide empirical evidence on A from samples representing the general population, no paper exists that directly measures A for taxpayers in the top bracket. There are reasons to expect top earners to differ from the general population in terms of misperception shape: the gap between the average and the marginal rate is largest at the top, which strengthens the effect of ironing, whereas the top statutory rate is a salient and widely discussed number, which strengthens over-perception.

Identified set. Based on the argument above, the set of pairs $\theta = (\bar{\tau}, A)^t$ supported by the data, i.e., the identified set for θ, is derived as follows. Suppose that reported incomes are observed with topcoding at a known cap $z_c > z^*$, so that the mean income $z_m(z^*)$ is bounded but not identified, and let $[\pi_L, \pi_U]$ denote the resulting range of the tail parameter π. Under appropriate conditions on the data and the behavioral parameters, the identified set for θ can be expressed as

$$\Theta_I = \left\{ (\bar{\tau}, A)^t \in \Theta : \; A_L \leq A \leq A_U, \; \frac{1}{1 + \bar{\zeta}_{c,r} \pi_U \kappa(A)} \leq \bar{\tau} \leq \frac{1}{1 + \bar{\zeta}_{c,r} \pi_L \kappa(A)} \right\}, \tag{3}$$

[3]Here I fix ω at a known value ω_0, for which salience studies provide estimates (e.g., Chetty et al., 2009; Taubinsky and Rees-Jones, 2018; Lardeux, 2023), although these come from different contexts.

where $[A_L, A_U]$ is a maintained range for the misperception shape. The formal set-up of the moment restrictions and the derivation of Θ_I are given in Section 3.3.

The set Θ_I has specific geometric features. First, it is not convex. Its upper boundary, $A \mapsto 1/(1+\bar{\zeta}_{c,r} \cdot \pi_L \cdot \kappa(A))$, is the reciprocal of a positive affine function of A and hence strictly convex for $\omega_0 < 1$, so the region below it is not convex. Second, the set is the intersection of four sets defined by smooth inequalities, so that its corners are outward kinks, and the slopes of the two inequalities on $\bar{\tau}$ depend on the data through π_L and π_U. These points suggest an alternative approach from the existing one which utilize the convexity of identified set. The support function approach (e.g., Beresteanu and Molinari (2008), Kaido and Santos (2014)) characterizes the convex hull of Θ_I, which contains pairs $(\bar{\tau}, A)$ that no admissible value of π generates; hence the dependence of the optimal rate on the misperception shape, which is the object of this exercise, cannot be inferred from this approach.

3. Modeling framework

3.1. **Problem settings.** Let $(\mathcal{X}, \mathcal{B}_\mathcal{X}, P)$ denote a probability space, where $\mathcal{X}$ is a Borel subset of a Euclidean space with its Borel σ-algebra $\mathcal{B}_\mathcal{X}$, and P belongs to a family of probability measures $\mathcal{P}$ on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. Let $\Theta \subset \mathbb{R}^{d_\theta}$ with $d_\theta < \infty$ be a compact set. Consider a moment function $m : \Theta \times \mathcal{X} \mapsto \mathbb{R}^{d_m}$, which is assumed to belong to a family $\mathcal{M}$ of measurable functions. Given a moment function m and a probability measure P, one obtains a moment restriction $\Phi : \Theta \times \mathcal{P} \mapsto \mathbb{R}^{d_q}$ defined as $\Phi(\theta, P) := q\left(\int_\mathcal{X} m(\theta, \cdot)\, dP(\cdot)\right)$, where $q : \mathbb{R}^{d_m} \mapsto \mathbb{R}^{d_q}$ is some known function. I split the restrictions into equality and inequality types: let $\mathcal{J}_1$ and $\mathcal{J}_2$ be disjoint sets with $\mathcal{J}_1 \cup \mathcal{J}_2 = \{1, \cdots, d_q\}$, and write $\Phi_{\mathcal{J}_k}(\theta, P) := (\Phi_j(\theta, P))_{j \in \mathcal{J}_k}$ for $k = 1, 2$.

My interest is the "identified set", the set of parameters satisfying the moment restrictions,

$$\Theta_I = \Theta_I(\Phi, P) := \left\{\theta \in \Theta : \Phi_{\mathcal{J}_1}(\theta, P) = 0 \in \mathbb{R}^{|\mathcal{J}_1|},\ \Phi_{\mathcal{J}_2}(\theta, P) \leq 0 \in \mathbb{R}^{|\mathcal{J}_2|}\right\}. \tag{4}$$

This formulation covers moment equality models ($\mathcal{J}_2 = \emptyset$), moment inequality models ($\mathcal{J}_1 = \emptyset$) and their combinations, and q allows Φ to be a nonlinear function of the moments. In Section 4 I derive the asymptotic efficiency bound for Θ_I defined in (4). Based on the result, in Section 5 I consider the set-valued estimator T_n based on criterion function $Q : \Theta \mapsto \mathbb{R}$, where T_n is defined as the contour set or level set of the sample analogue Q_n. The set Θ_I, which may contain more than one point, is assumed to belong to a certain class $\mathcal{K}^*_\Theta$ of sets specified in detail later. A key observation here is that properties of Θ_I depend on the moment restriction Φ and a probability measure P.

One of my objectives is to derive and characterize the efficiency bound for Θ_I among a certain class of set-valued estimators T_n. In the point identified case, one needs proper restrictions on the model $\mathcal{P}$ and the class of estimators. In particular, the model $\mathcal{P}$ is assumed to be "smooth" often measured in the L^2-norm of its density (the so-called "differentiable in quadratic mean"), while the class of estimators is assumed to behave

the same way in neighborhoods of parameters ("locally regular"); see Bickel et al. (1998) for textbook treatment.

The key difficulties in my context are (1) how these conditions on $\mathcal{P}$ and the class of estimators in the point identified case are transferred to the set-identified case and (2) what conditions on $\mathcal{K}_\Theta^*$ are appropriate. For (2), I set up $\mathcal{K}_\Theta^*$ to allow non-convex sets there; this widens the scope of Kaido and Santos (2014), where the target class $\mathcal{K}_\Theta^*$ is restricted to convex sets. I study these conditions in detail in the set-identified case.

3.2. **Class of compact sets $\mathcal{K}_\Theta^*$.** I specify the target class of sets $\mathcal{K}_\Theta^*$ to which the identified set Θ_I is assumed to belong. The point here is that the class $\mathcal{K}_\Theta^*$ should be nice enough such that analysis to derive efficiency bound and limiting distribution of estimators in point-identified case can be naturally extended to set case, while at the same time the class should be flexible enough for application.

Let $\mathcal{K}$ be the family of all non-empty compact sets in $\mathbb{R}^{d_\theta}$. The space $(\mathcal{K}, d_H)$, where $d_H : \mathcal{K} \times \mathcal{K} \mapsto \mathbb{R}_+$ is the Hausdorff metric, is then a complete separable metric space[4]. The target class $\mathcal{K}_\Theta^*$ is assumed to be $\mathcal{K}_\Theta^* \subset \mathcal{K}_\Theta := \{K \in \mathcal{K} : K \subset \Theta\}$, whose characteristic is specified as follows.

First, let $\mathcal{K}_1 \subset \mathcal{K}_\Theta$ be such that

$$\mathcal{K}_1 := \{K \in \mathcal{K}_\Theta \text{ in } \mathbb{R}^{d_\theta} : |K| \leq N_1 < \infty\}.$$

The class $\mathcal{K}_1$ consists of finite sets with at most N_1 points. Next, define a class $\mathcal{K}_2 \subset \mathcal{K}_\Theta$ as

$$\mathcal{K}_2 := \Big\{K \in \mathcal{K}_\Theta : d_\theta = 1,\ K = \bigcup_{i=1}^{N} K_i \text{ for some } N \leq N_2, \text{ where } K_1, \cdots, K_N \text{ are disjoint compact intervals with non-empty interior}\Big\},$$

for some $N_2 < \infty$. The class $\mathcal{K}_2$ consists of finite unions of at most N_2 disjoint compact intervals.

Now I define the last compact class $\mathcal{K}_3 \subset \mathcal{K}_\Theta$. The class is designed so that its members have a piecewise smooth boundary and may have a dimension lower than d_θ. A geometric notion that captures these properties is manifold. To formalize this intuition, I follow the definitions in Lee (2012). Let $\mathbb{H}^k := \{x = (x^1, \ldots, x^k) \in \mathbb{R}^k : x^k \geq 0\}$ denote the closed upper half-space and $\overline{\mathbb{R}}_+^k := \{x \in \mathbb{R}^k : x^1 \geq 0, \ldots, x^k \geq 0\}$ the closed positive orthant, which is the model for corners.

Definition 3.1 (Smooth manifold with corners)**.** Let $k \geq 1$ be an integer.

(a) A *topological k-manifold with boundary* is a second-countable Hausdorff space M in which every point has a neighborhood homeomorphic either to an open subset of $\mathbb{R}^k$ or to a relatively open subset of $\mathbb{H}^k$.

[4]More generally, the space of non-empty compact sets is complete with the Hausdorff metric if the underlying space is locally compact or complete (Schneider (2013), Molchanov (2017)). Separability also holds if the underlying space is a separable Banach space (Giné et al. (1983)).

(b) Let M be a topological k-manifold with boundary. An *interior chart* for M is a pair (U, φ), where $U \subset M$ is open and φ is a homeomorphism from U onto an open subset of $\mathbb{R}^k$. A *chart with corners* for M is a pair (U, φ), where $U \subset M$ is open and φ is a homeomorphism from U onto a relatively open subset of $\overline{\mathbb{R}}^k_+$.

(c) Two such charts (U, φ) and (V, ψ) are *smoothly compatible* if the transition map $\varphi \circ \psi^{-1} : \psi(U \cap V) \mapsto \varphi(U \cap V)$ is smooth, i.e., it admits a C^∞ extension to an open neighborhood of each point of $\psi(U \cap V)$.

(d) A *smooth structure with corners* on M is a maximal collection of smoothly compatible interior charts and charts with corners whose domains cover M. A topological k-manifold with boundary together with a smooth structure with corners is called a *smooth k-manifold with corners*.

Parts (a) and (b)–(d) are the definitions in Lee (2012, pp. 25 and 415), respectively. Note that every smooth manifold with or without boundary is also a smooth manifold with corners and that smooth maps, tangent spaces and differentials on smooth manifolds with corners are defined through these charts in exactly the same way as on smooth manifolds with boundary (Lee, 2012, p. 417). Accordingly, following Lee (2012, p. 85), a map $\iota : M \mapsto \mathbb{R}^{d_\theta}$ from a smooth manifold with corners M is a *smooth embedding* if it is a smooth immersion (i.e., its differential is injective at every $x \in M$) and a homeomorphism onto its image $\iota(M)$ with the subspace topology.

Now I define a target set based on Definition 3.1 in the following manner.

Definition 3.2 (Embedded submanifold with corners)**.** Let $k \in \{1, \dots, d_\theta\}$. A subset $K \subset \mathbb{R}^{d_\theta}$ is a *k-dimensional embedded submanifold with corners* of $\mathbb{R}^{d_\theta}$ if K, endowed with the subspace topology, is a topological k-manifold with boundary equipped with a smooth structure with corners such that the inclusion map $\iota_K : K \hookrightarrow \mathbb{R}^{d_\theta}$ is a smooth embedding.

Definition 3.2 is the analogue of an embedded submanifold with boundary, with charts with corners in place of boundary charts (Lee, 2012, p. 120). Note that the requirement that ι_K be a smooth embedding is important[5]. I now define

$$\mathcal{K}_3 := \big\{K \in \mathcal{K}_\Theta : K \text{ is a compact } k\text{-dimensional embedded submanifold with corners of } \mathbb{R}^{d_\theta} \text{ for some } k \in \{1, \dots, d_\theta\}\big\}.$$

The class $\mathcal{K}_3$ includes full-dimensional sets ($k = d_\theta$), such as a closed ball or a polytope with non-empty interior, as well as lower-dimensional sets ($k < d_\theta$), such as a sphere or a curve segment in $\mathbb{R}^3$[6].

For the target class $\mathcal{K}^*_\Theta$, I focus on the union of each compact class described above.

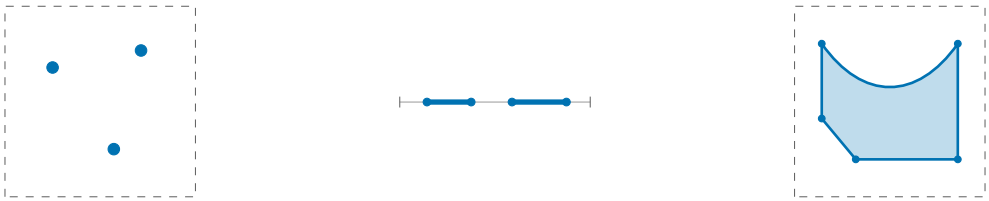
[5]Without this requirement, Definition 3.1 only requires that *some* smooth structure with corners exists on K, which places no restriction on how K sits in $\mathbb{R}^{d_\theta}$. For example, the Koch snowflake curve in $\mathbb{R}^2$ is homeomorphic to the unit circle and hence admits a smooth structure, although it has no tangent line at any point.

[6]Note that throughout the paper, ∂K and K^o denote the *topological* boundary and interior of K as a subset of $\mathbb{R}^{d_\theta}$, which are in general different from the manifold boundary and interior of K (Lee, 2012, p. 26).

Assumption 3.3 (Class of compact sets $\mathcal{K}_\Theta^*$)**.**

$$\Theta_I \in \mathcal{K}_\Theta^* = \begin{cases} \mathcal{K}_1 \cup \mathcal{K}_2 & \text{if } d_\theta = 1, \\ \mathcal{K}_1 \cup \mathcal{K}_3 & \text{if } d_\theta \geq 2. \end{cases}$$

The class $\mathcal{K}_1$ roughly captures the point identified case and the case with multiple distinct singletons. The class $\mathcal{K}_2$ corresponds to the (multiple of) interval identified set. The class $\mathcal{K}_3$ reflects the case where the identified set has piecewise smooth boundary, which is often observed in moment inequality models.



(a) $\mathcal{K}_1$ (b) $\mathcal{K}_2$ (c) $\mathcal{K}_3$

FIGURE 1. Illustrative members of the classes $\mathcal{K}_1$, $\mathcal{K}_2$ and $\mathcal{K}_3$.

Remark 3.4. Note that the classes $\mathcal{K}_1$ and $\mathcal{K}_2$ can be seen as special cases of $\mathcal{K}_3$ if $k = 0$ is allowed in Definitions 3.1 and 3.2. A compact 0-dimensional manifold is a finite set, so every $K \in \mathcal{K}_1$ is a compact 0-dimensional embedded submanifold of $\mathbb{R}^{d_\theta}$; and for $d_\theta = 1$, every $K \in \mathcal{K}_2$ is a compact 1-dimensional embedded submanifold with corners of $\mathbb{R}$, since smooth 1-manifolds with corners are the same as smooth 1-manifolds with boundary (Lee, 2012, p. 416). I keep the three classes separate for ease of exposition; the bounds N_1 and N_2 make $\mathcal{K}_1$ and $\mathcal{K}_2$ smaller than the corresponding parts of $\mathcal{K}_3$.

3.3. **Example revisited: optimal top tax rate.** I here cast the example of Section 2 in the framework above. Recall from (2) that, for a known attention weight ω_0, the optimal top rate is $\bar{\tau} = 1/(1 + \bar{\zeta}_{c,r} \cdot \pi \cdot \kappa(A))$ with $\kappa(A) = \omega_0 + (1 - \omega_0)A$, and that the parameter is $\theta = (\bar{\tau}, A)^t$, $d_\theta = 2$.

Data. Let $Z \geq 0$ denote reported income, which is not observed when it exceeds a known cap $z_c > z^*$. The observed variable is $X = (\tilde{Z}, D)$ with $\tilde{Z} := Z \wedge z_c$ and $D := \mathbb{1}\{Z \geq z_c\}$, and $\{X_i\}_{i\in[n]}$ is a random sample from P on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. Define the identified quantities

$$\begin{aligned} S &:= P(\tilde{Z} \geq z^*), \\ M_L &:= \mathbb{E}_P\left[\tilde{Z} \cdot \mathbb{1}\{\tilde{Z} \geq z^*\}\right], \\ M_U &:= \mathbb{E}_P\left[\left(\tilde{Z} \cdot (1 - D) + \bar{z} \cdot D\right) \cdot \mathbb{1}\{\tilde{Z} \geq z^*\}\right], \end{aligned}$$

where $\bar{z}$ is given in Assumption 3.5(ii) below. M_L replaces each topcoded income by the cap, and M_U by the largest admissible mean income above the cap.

Assumption 3.5 (Top tax rate)**.** (i). $A \in [A_L, A_U]$ for a known range with $0 < A_L < A_U$. (ii). $\mathbb{E}[Z \mid Z \geq z_c] \leq \bar{z}$ for a known $\bar{z} > z_c$. (iii). The attention weight $\omega_0 \in (0,1)$ and the elasticity $\bar{\zeta}_{c,r} > 0$ are known. (iv). $P(\tilde{Z} > z^*) > 0$ and $P(D = 1) > 0$.

Assumption 3.5 (i) is the only restriction on misperception. For example, if every taxpayer in the bracket either irons or responds at least as strongly as a correctly perceiving taxpayer, then A_L can be taken at the ironing value $1/\pi$ evaluated at the largest admissible π, while $A_U > 1$ bounds the degree of over-perception. Assumption 3.5 (ii) is used only on the topcoded cell, which implies that the mean income above z^* is finite so that $\pi > 1$. It bounds the mean income above the cap, not the support of Z, and hence unbounded Pareto tails are allowed. Note that the expectation is taken with respect to the distribution of Z, which is well defined since $P(D = 1) > 0$ by Assumption 3.5 (iv). Assumption 3.5 (iii) treats the elasticity as known for simplicity[7]. Assumption 3.5 (iv) rules out a degenerate bracket and ensures that π is not point identified.

Parameter space. Let $\Theta := [\tau_L^*, \tau_U^*] \times [A_L^*, A_U^*]$ with $0 < \tau_L^* < \tau_U^* < 1$ and $0 < A_L^* < A_L < A_U < A_U^*$, where τ_L^* and τ_U^* are chosen so that Θ_I below lies in the interior of Θ. Note that $\kappa(A) > 0$ on Θ. Then inverting (2) gives the tail parameter that supports θ,

$$\pi(\theta) := \frac{1}{\bar{\zeta}_{c,r} \cdot \kappa(A)} \left(\frac{1}{\bar{\tau}} - 1 \right), \tag{5}$$

which is C^∞ on Θ, with $\partial \pi(\theta)/\partial \bar{\tau} = -1/(\bar{\zeta}_{c,r} \cdot \kappa(A) \cdot \bar{\tau}^2) < 0$. Note that the map $\theta \mapsto (\pi(\theta), A)$ is a change of coordinates rather than a projection, since the coordinate A is retained.

Moment restrictions. Let q be the identity and $m = (m_1, \cdots, m_4) : \Theta \times \mathcal{X} \mapsto \mathbb{R}^4$ be given by

$$\begin{aligned}
m_1(\theta, X) &= \left[(1 - \pi(\theta)) \left(\tilde{Z} \cdot (1 - D) + \bar{z} \cdot D \right) + \pi(\theta) \cdot z^* \right] \cdot \mathbb{1}\{\tilde{Z} \geq z^*\}, \\
m_2(\theta, X) &= \left[(\pi(\theta) - 1)\, \tilde{Z} - \pi(\theta) \cdot z^* \right] \cdot \mathbb{1}\{\tilde{Z} \geq z^*\}, \\
m_3(\theta, X) &= A_L - A, \quad m_4(\theta, X) = A - A_U,
\end{aligned}$$

so that

$$\Phi_1(\theta, P) = (1 - \pi(\theta))\, M_U + \pi(\theta) z^* S, \quad \Phi_2(\theta, P) = (\pi(\theta) - 1)\, M_L - \pi(\theta) z^* S.$$

Both stochastic constraints are affine in the moments (S, M_L, M_U) with coefficients that depend on θ, and they are defined on all of Θ; the restrictions $\Phi_3(\theta)$ and $\Phi_4(\theta)$ do not depend on P.

Under the setting above, the identified set Θ_I is expressed as in (3). The formal statement is the following.

[7] If $\bar{\zeta}_{c,r}$ is estimated, it enters Φ below through q. I abstract from this here and leave it for future research.

Proposition 3.6 (Identified set for the top tax rate). *Suppose Assumption 3.5 holds. Let $\mathcal{Q}(P)$ be the set of distributions Q of Z on $[0,\infty)$ with $\mathbb{E}_Q[Z \mid Z \geq z_c] \leq \bar{z}$ under which $(Z \wedge z_c, \mathbb{1}\{Z \geq z_c\}) \sim P$. For $Q \in \mathcal{Q}(P)$, let $z_{m,Q} := \mathbb{E}_Q[Z \mid Z \geq z^*]$ and $\pi_Q := z_{m,Q}/(z_{m,Q} - z^*)$. Finally define*

$$\Theta_I := \{\theta \in \Theta : A_L \leq A \leq A_U,\ \pi(\theta) = \pi_Q \text{ for some } Q \in \mathcal{Q}(P)\}.$$

Then the following statements hold.

(a) $\{\pi_Q : Q \in \mathcal{Q}(P)\} = [\pi_L, \pi_U]$ *with*

$$\pi_L := \frac{M_U/S}{M_U/S - z^*}, \quad \pi_U := \frac{M_L/S}{M_L/S - z^*}, \quad 1 < \pi_L < \pi_U < \infty.$$

(b) $\Theta_I = \{\theta \in \Theta : \Phi_j(\theta, P) \leq 0,\ j = 1, \cdots, 4\}$, *and Θ_I equals the set in (3).*

Proof. See Appendix B.1. □

4. Efficiency bounds for identified set

My purpose here is to derive the lower bound of the asymptotic risk

$$\liminf_{n\to\infty} R_{P,l}(T_n; A) := \liminf_{n\to\infty} \int_{\mathcal{X}^n} l \circ d_H\left(T_n(x^n), A\right)\ dP(x^n), \tag{6}$$

where $l \circ d_H : \mathcal{K}_\Theta \times \mathcal{K}_\Theta \mapsto \mathbb{R}_+$ denotes a loss function which is some composition of the Hausdorff distance d_H, a set $A \in \mathcal{K}_\Theta$ is of interest, and T_n is a (possibly scaled) $\mathcal{K}_\Theta$-valued estimator. The case $A = \Theta_I$ with $d_H(T_n, \Theta_I) = o_P(1)$ (and hence $R_{P,l}(T_n; \Theta_I) = o(1)$ for well-behaved l such as continuous and bounded ones) is a minimal consistency result obtained in the literature. I aim to go further to characterize the efficiency of T_n by focusing on its behavior in the local neighborhood of Θ_I. Note that, since the identified set is determined by the distribution of the data, the relevant set under a local data-generating distribution P is the identified set under P, which I denote by $A = \Theta_I(P)$, rather than a fixed set shifted by a local parameter; I make this point precise in Section 4.1.

To analyze the asymptotic properties of the risk in such a situation, I start my analysis with the case where identified set is defined via a likelihood function. In Section 4.1 I first establish the asymptotic representation for set-valued statistic T_n via the limit experiment framework that is tailored to my context. Based on this result, in Section 4.2 I derive the asymptotic lower bound. In the process of deriving the results, I discuss how the key notions used in point identified case are naturally extended to the set-identified context. In Section 4.3 I extend these results to identified sets defined by the moment restrictions in (4). Note that Sections 4.1 and 4.2 focus on identified sets that are finite or lower-dimensional smooth sets, for which the identified set under a local deviation of the likelihood moves at the rate $\sqrt{n}$; identified sets with non-empty interior such as intervals are handled in Section 4.3, where the local deviation is taken with respect to the distribution of the data.

From now on, the Borel σ-algebra $\mathcal{B}_{\mathcal{K}_\Theta}$ generated by the Hausdorff metric d_H is defined on $\mathcal{K}_\Theta$, where recall $\mathcal{K}_\Theta := \{K \in \mathcal{K} : K \subset \Theta\}$ and $\mathcal{K}$ is the family of all non-empty compact sets in $\mathbb{R}^{d_\theta}$. Similarly $\mathcal{B}_\mathcal{K}$ denotes the Borel σ-field on $\mathcal{K}$. Since the space $(\mathcal{K}, d_H)$ is complete and separable, so is $(\mathcal{K}_\Theta, d_H)$. The measurability of "set-valued" random variables $T_n : \mathcal{X}^n \mapsto \mathcal{K}_\Theta$ is hence defined accordingly with respect to the measurable space $(\mathcal{K}_\Theta, \mathcal{B}_{\mathcal{K}_\Theta})$[8].

4.1. **Limit experiment for identified set.** Consider a parametric likelihood function $p_\theta = dP_\theta/d\mu$ on the sample space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ with a σ-finite measure μ. The model is indexed by a finite dimensional parameter $\theta \in \Theta^o \subset \mathbb{R}^{d_\theta}$. For simplicity, I consider the situation where potentially multiple parameters induce observational equivalence, i.e., the equivalence $p_{\theta_1} = p_{\theta_2}$ holds μ-a.e. even when $\theta_1 \neq \theta_2$. Following Chen et al. (2018), given a true density $p_0 = dP_0/d\mu$, the parameters that maximize the likelihood are those with $\mathrm{KL}(p_0|p_\theta) := \int_\mathcal{X} \log(p_0/p_\theta) \cdot p_0 \; d\mu = 0$, and their set is Θ_I. I want to estimate, as well as possible, the set Θ_I of parameters that are consistent with the model $\{p_\theta : \theta \in \Theta^o \subset \mathbb{R}^{d_\theta}\}$ and the available data $\{X_i\}_{i\in[n]}$. More generally, for a probability measure P on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ whose density $p = dP/d\mu$ belongs to the model, I denote by

$$\Theta_I(P) := \{\theta \in \Theta^o : p_\theta = p \;\; \mu\text{-a.e.}\} \tag{7}$$

the identified set under P (defined over Θ^o, on which the model is defined, unlike (4)), so that $\Theta_I = \Theta_I(P_0)$; I write $\Theta_I(P)$ to make the dependence on P explicit. The set $\Theta_I(P)$ plays a key role below, as the identified set under a local deviation of the data-generating distribution is the natural target of set estimators.

To set up the main idea, suppose that the likelihood p_θ satisfy the "*differentiable in quadratic mean* (DQM)" condition—recall that a set of probability densities (or "experiment") $\{p_\theta : \theta \in \Theta^o \subset \mathbb{R}^{d_\theta}\}$ is said to satisfy the DQM at $\theta_0 \in \Theta^o$ if there exists a measurable function $\dot{\ell}_{\theta_0}$ such that

$$\int_\mathcal{X} \left(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2}\dot{\ell}_{\theta_0} h \sqrt{p_{\theta_0}} \right)^2 d\mu = o(\|h\|^2), \tag{8}$$

as $h \to 0$ in $\mathbb{R}^{d_\theta}$. The measurable function $\dot{\ell}_\theta$ is the score at θ; it can be shown that $P_\theta \dot{\ell}_\theta = 0$ and that there exists the Fisher information $I_\theta = P_\theta \dot{\ell}_\theta^t \dot{\ell}_\theta$ (e.g. van der Vaart (1998)). Two features can be observed in the point identified case. First, the score at an identified point θ_0 is uniquely determined as $\dot{\ell}_{\theta_0}$ regardless of the direction of h, which plays key roles to derive (asymptotic) efficiency bound. Second, the non-singularity of the Fisher information $I_\theta > 0$ is typically assumed; this is a type of regularity condition which holds for standard models. As shown below, in general these features no longer hold in set identified cases.

When the object of interest is the set Θ_I that yields observational equivalence, a natural candidate for the local deviation of Θ_I would be the Minkowski sum $\Theta_I \oplus h/\sqrt{n} = \{\theta + h/\sqrt{n} : \theta \in \Theta_I\}$. This set is, however, not the identified set under the

[8]The topology on the family of non-empty compact sets induced by the Hausdorff metric is equivalent to *Myopic topology* if the underlying space is a separable metric space; see Appendix A.2.

locally deviated model in general: if the data are generated by $P_{\theta+h/\sqrt{n}}$ with $\theta \in \Theta_I$, the set of parameters that are consistent with the model and the data is $\Theta_I(P_{\theta+h/\sqrt{n}})$ in (7), and different points of Θ_I can move in different directions[9]. Accordingly, I take the identified set $\Theta_I(P_{\theta+h/\sqrt{n}})$ under the locally deviated model as the local deviation of Θ_I in what follows; this is where the difference from the point identified scenario arises and some care is required.

The information of local asymptotics, i.e., $\lim_{h\to 0} p_{\theta+h}/p_\theta$, is the key quantity, but this quantity in my setting depends on the location of $\theta \in \Theta_I$ and on the direction of $h \in \mathbb{R}^{d_\theta}$. More precisely, suppose that the model $\{p_\theta : \theta \in \Theta^o \subset \mathbb{R}^{d_\theta}\}$ satisfies the DQM in (8) at $\theta \in \Theta_I$. Since $p_{\theta'} = p_\theta$ μ-a.e. for any $\theta' \in \Theta_I$, the score $\dot{\ell}_\theta h$ vanishes for any direction h along which Θ_I extends from θ, i.e., for h in the tangent cone $T(\theta)$ defined in (11) below, so that, unless θ is an isolated point of Θ_I, the Fisher information I_θ is necessarily singular and only the directions leaving Θ_I can carry information. When Θ_I has a corner or non-empty interior, the DQM in (8) is even too restrictive: at a boundary point the score $\dot{\ell}_\theta h$ is linear in h and vanishes on an open cone of inward directions, so that it vanishes for all h, whereas the directions leaving Θ_I do carry information[10]. This observation implies that each direction expressed by h could provide different information, i.e. the directional score $\dot{\ell}_{\theta,h}h$ and its variance $P_0(\dot{\ell}_{\theta,h}h)^2$ depend on the location of $\theta \in \Theta_I$ and direction h, which is a sharp contrast to the point identified case with Fréchet differentiability of the model $\{p_\theta : \theta \in \Theta^o \subset \mathbb{R}^{d_\theta}\}$ where any local deviation leads to a single I_θ.

I now formalize the intuition built above. Given that the model has different scores depending on the location of parameters and direction, it is not appropriate to characterize scores via its Fréchet derivative because it requires uniqueness regardless of how its limit is taken (i.e. invariant of direction). I instead use the following notion.

Definition 4.1 (Directional score)**.** A model $\{p_\theta : \theta \in \Theta^o \subset \mathbb{R}^{d_\theta}\}$ is said to have the directional score at $\theta \in \Theta^o$ for $h \in \mathbb{R}^{d_\theta}$, denoted by $\dot{\ell}_{\theta,h}$, if there exists a non-empty set $H_\theta \subset \mathbb{R}^{d_\theta}$ such that $\dot{\ell}_{\theta,h}h \neq 0$ for any $h \in H_\theta$ and

$$\int_{\mathcal{X}} \left(\frac{\sqrt{p_{\theta+t_n h_n}} - \sqrt{p_\theta}}{t_n} - \frac{1}{2}\dot{\ell}_{\theta,h} h \sqrt{p_\theta} \right)^2 d\mu = o(1), \tag{9}$$

for any $\{h_n\}_{n\in\mathbb{N}} \subset H_\theta$ and any $\{t_n\}_{n\in\mathbb{N}} \subset \mathbb{R}_+$ with $h_n \to h \in H_\theta$ and $t_n \to 0$ as $n \to \infty$.

The crucial difference from (8) is that the score $\dot{\ell}_{\theta,h}$ depends on h. Note, however, that (9) determines only the real-valued function $\dot{\ell}_{\theta,h}h$, the score along h, and not $\dot{\ell}_{\theta,h}$ itself; adding to $\dot{\ell}_{\theta,h}$ any square-integrable $b : \mathcal{X} \mapsto \mathbb{R}^{d_\theta}$ with $b(x)^t h = 0$ for μ-a.e. x

[9]For example, consider $X \sim N(g(\theta), 1)$ with $g(\theta) = (\theta - \theta_1)(\theta - \theta_2)$ and $\theta_1 \neq \theta_2$, where $\mathbb{E}X = 0$ under the true distribution, so that $\Theta_I = \{\theta_1, \theta_2\}$. Since $g(\theta_1 + t) = g(\theta_2 - t)$ for any $t \in \mathbb{R}$, it holds that $\Theta_I(P_{\theta_1+h/\sqrt{n}}) = \{\theta_1 + h/\sqrt{n}, \theta_2 - h/\sqrt{n}\}$, whereas $\Theta_I \oplus h/\sqrt{n} = \{\theta_1 + h/\sqrt{n}, \theta_2 + h/\sqrt{n}\}$; moreover, the two local parameters $\theta_1 + h/\sqrt{n}$ and $\theta_2 - h/\sqrt{n}$ induce the same distribution of the data but different Minkowski sums, so that the latter cannot be recovered from the data even in the limit.

[10]This is because the score $\dot{\ell}_\theta$ in (8) is the Fréchet derivative of the map $\phi : \Theta \to L^2(\mu)$ with $\phi(\theta) = \sqrt{p_\theta}$, which is linear in the direction h.

leaves $\dot{\ell}_{\theta,h}h$ unchanged. Accordingly, for any $\theta \in \Theta_I$ the information in direction h is measured by the scalar

$$\sigma_\theta(h) := P_\theta \left(\dot{\ell}_{\theta,h}h\right)^2, \tag{10}$$

which equals $P_0(\dot{\ell}_{\theta,h}h)^2$ for $\theta \in \Theta_I$, as $p_\theta = p_0$ μ-a.e., and depends on the direction h; in terms of $I_{\theta,h} := P_\theta \dot{\ell}^t_{\theta,h}\dot{\ell}_{\theta,h}$ it holds that $\sigma_\theta(h) = h^t I_{\theta,h} h$, which is the only part of $I_{\theta,h}$ determined by the model. In this sense, both $\dot{\ell}_{\theta,h}h$ and $\sigma_\theta(h)$ can be viewed as maps on $\Theta^o \times \mathbb{R}^{d_\theta}$, capturing interaction of location of θ and direction h. Note that this definition implies $\dot{\ell}_{\theta,h}h$ being continuous and positively homogeneous of degree one in $h \in H_\theta$. In short, Definition 4.1 says that the map $\theta \mapsto \sqrt{p_\theta}$ is directionally differentiable at $\theta \in \Theta^o$ tangentially to H_θ in the sense of Hadamard; see Shapiro (1990)[11]. To describe H_θ, I use the following geometric quantities of Θ_I for $\theta \in \Theta_I$;

$$\begin{aligned} T(\theta) &:= \{h \in \mathbb{R}^{d_\theta} : \theta + t_n h_n \in \Theta_I \ \text{ for some } t_n \searrow 0 \text{ and } h_n \to h \ \}, \\ N(\theta) &:= \{v \in \mathbb{R}^{d_\theta} : v^t h \leq 0 \ \text{ for all } h \in T(\theta)\}. \end{aligned} \tag{11}$$

$T(\theta)$ is the tangent cone of Θ_I at θ and $N(\theta)$ is the regular normal cone (Rockafellar and Wets, 1998, Definitions 6.1 and 6.3 and Proposition 6.5). If θ is an interior point of Θ_I, then $T(\theta) = \mathbb{R}^{d_\theta}$ and $N(\theta) = \{0\}$. These notions of directional scores and geometric characteristics are used below to describe the limit of the locally deviated models and the distance between their identified sets.

The main result in this section builds on the local behavior of the likelihood described in Definition 4.1. Consider a model $\{p^n_\theta := \prod_{i=1}^n p_\theta : \theta \ \in \Theta^o \subset \mathbb{R}^{d_\theta}\}$ which yields the set of observationally equivalent parameters Θ_I. I focus on estimators $T_n : \mathcal{X}^n \mapsto \mathcal{K}_\Theta$ in the locally deviated models

$$\left\{P^n_{\theta+h/\sqrt{n}} : \theta \in \Theta_I, \ h \in \mathbb{R}^{d_\theta}\right\}, \tag{12}$$

where $P^n_{\theta+h/\sqrt{n}}$ denotes the n-fold product of $P_{\theta+h/\sqrt{n}}$, i.e., the data are generated by the parameter $\theta + h/\sqrt{n}$ for some $\theta \in \Theta_I$ and $h \in \mathbb{R}^{d_\theta}$. Note that the local deviation is indexed by the pair (θ, h) rather than by h alone, since the information of local asymptotics depends on both the location of $\theta \in \Theta_I$ and the direction h, as seen above. The quality of T_n is measured by the (scaled) Hausdorff metric $\sqrt{n} \cdot d_H(T_n, \Theta_I(P_{\theta+h/\sqrt{n}}))$ between T_n and the identified set under the locally deviated model. Note here that the object of interest can be viewed as an $\mathbb{R}_+$-valued random variable $\sqrt{n} \cdot d_H(T_n, \Theta_I(P_{\theta+h/\sqrt{n}})) : \mathcal{X}^n \mapsto \mathbb{R}_+$. I show that its law can be asymptotically equivalent to that of a functional of a statistic in a Gaussian limit experiment. To this end, I impose the following conditions on the local behavior of the identified set $\Theta_I(P_{\theta+h/\sqrt{n}})$.

[11]The Hadamard directional differentiability is often employed to analyze inference of parameters which can be characterized via directionally differentiable maps and hence a standard Delta method based on full differentiability cannot be applied (e.g. Fang and Santos (2019)).

Let $\mathcal{C}(\Theta_I)$ denote the space of continuous maps $g : \Theta_I \mapsto \mathbb{R}^{d_\theta}$ equipped with the supremum norm $\|g\|_\infty := \sup_{\vartheta\in\Theta_I} \|g(\vartheta)\|$. For $g \in \mathcal{C}(\Theta_I)$ and $t > 0$, define

$$\Theta_I \oplus_t g := \{\vartheta + t\, g(\vartheta) : \vartheta \in \Theta_I\}, \tag{13}$$

which moves each point $\vartheta \in \Theta_I$ by $t\, g(\vartheta)$; if $g \equiv h$ is constant, $\Theta_I \oplus_t g = \Theta_I \oplus th$ is the Minkowski sum.

Assumption 4.2 (Local deviation of the identified set)**.** The model $\{p_\theta : \theta \in \Theta^o \subset \mathbb{R}^{d_\theta}\}$ and the identified set Θ_I satisfy the following conditions;

(i). Θ_I is either a finite set in $\mathcal{K}_1$, or a compact k-dimensional embedded smooth submanifold of $\mathbb{R}^{d_\theta}$ without boundary with $1 \le k < d_\theta$,

(ii). for every $\theta \in \Theta_I$ and $h \in \mathbb{R}^{d_\theta}$, (9) holds with $h_n = h$ and $t_n = 1/\sqrt{n}$, where $\dot\ell_{\theta,h} h := 0$ for $h \notin H_\theta$,

(iii). for every $\theta \in \Theta_I$ and $h \in \mathbb{R}^{d_\theta}$, there exists $g_{\theta,h} \in \mathcal{C}(\Theta_I)$ such that

$$\lim_{n\to\infty} \sqrt{n} \cdot d_H\left(\Theta_I(P_{\theta+h/\sqrt{n}}),\ \Theta_I \oplus_{1/\sqrt{n}} g_{\theta,h}\right) = 0. \tag{14}$$

Assumption 4.2 (i) specifies the class of identified sets treated in Sections 4.1 and 4.2; finite sets and smooth lower-dimensional sets such as spheres and closed curves are included. Assumption 4.2 (ii) only requires the expansion (9) along the path $\theta + h/\sqrt{n}$, and allows a zero score in the directions $h \notin H_\theta$, e.g. the directions along which Θ_I extends. Assumption 4.2 (iii) is the key condition: it states that the identified set under the locally deviated model moves at the rate $\sqrt{n}$, where each point $\vartheta \in \Theta_I$ is moved by $g_{\theta,h}(\vartheta)/\sqrt{n}$. Note that $g_{\theta,h}$ need not be constant, i.e. the identified set does not move as a rigid shift in general[12]. Assumption 4.2 (iii) excludes identified sets with non-empty interior arising from a likelihood, as the identified set under a local deviation across the boundary collapses[13]; such sets are typically defined via moment inequalities and are treated in Section 4.3.

I next describe the limit of the locally deviated models in (12). By Assumption 4.2 (ii), for any finite collection of pairs $(\theta_j, h_j) \in \Theta_I \times \mathbb{R}^{d_\theta}$, $j = 1, \cdots, J$, it holds that

$$\log \frac{dP^n_{\theta_j+h_j/\sqrt{n}}}{dP^n_0} = \mathbb{G}_n\left(\dot\ell_{\theta_j,h_j} h_j\right) - \frac{1}{2}\sigma_{\theta_j}(h_j) + o_{P_0}(1), \quad j = 1, \cdots, J, \tag{15}$$

where $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P_0)$ is the empirical process with $\mathbb{P}_n$ the empirical measure, and that $P^n_{\theta_j+h_j/\sqrt{n}}$ and P^n_0 are mutually contiguous; see Lemma B.1 in Appendix B.2. The expansion (15) indicates that the locally deviated model depends on (θ, h) only through the scalar score $\dot\ell_{\theta,h} h$, which leads to the following limit experiment. Let $\mathcal{G}$ be a centered Gaussian process indexed by $\{s \in L^2(P_0) : P_0 s = 0\}$ with covariance $\mathbb{E}\,\mathcal{G}(s)\mathcal{G}(s') = P_0(ss')$, defined on a probability space with probability measure G_0; the

[12] In the example of footnote 9, $g_{\theta_1,h}(\theta_1) = h$ and $g_{\theta_1,h}(\theta_2) = -h$.

[13] For example, let $Y \sim N(\kappa(\theta), 1)$ with $\kappa(\theta) = (\theta - b)_+ - (a - \theta)_+$ and $a < b$, so that $\Theta_I = [a, b]$; then $\Theta_I(P_{b+h/\sqrt{n}}) = \{b + h/\sqrt{n}\}$ for $h > 0$, which is at the Hausdorff distance $b - a$ from Θ_I although $P^n_{b+h/\sqrt{n}}$ and P^n_b are contiguous, so that no estimator can attain a finite risk for such a target.

scores $\dot{\ell}_{\theta,h}h$ belong to this index set, as $P_0\dot{\ell}_{\theta,h}h = 0$ by Lemma B.1[14]. For such s define the probability measure G_s by

$$\frac{dG_s}{dG_0} := \exp\left(\mathcal{G}(s) - \frac{1}{2}P_0 s^2\right). \tag{16}$$

The limit experiment is $\{G_s : s = \dot{\ell}_{\theta,h}h,\ \theta \in \Theta_I,\ h \in \mathbb{R}^{d_\theta}\}$ with the observation $\mathcal{G}$, and the locally deviated model $P^n_{\theta+h/\sqrt{n}}$ corresponds to G_s with $s = \dot{\ell}_{\theta,h}h$. Note that the parameter of the limit experiment is the score s rather than the pair (θ, h); two pairs (θ, h) and (θ', h') with the same score lead to the same limit.

To state the limit of the risk, I need a measure of the distance between two sets of the form (13). For $\Delta \in \mathcal{C}(\Theta_I)$, define

$$\|\Delta\|_\perp := \sup_{\vartheta\in\Theta_I} \sup\left\{\left|u^t\Delta(\vartheta)\right| : u \in N(\vartheta),\ \|u\| \le 1\right\}, \tag{17}$$

where $N(\vartheta)$ is the normal cone in (11). Under Assumption 4.2 (i), the tangent cone $T(\vartheta)$ in (11) is a linear subspace for every $\vartheta \in \Theta_I$ (the tangent space of Θ_I at ϑ, and $T(\vartheta) = \{0\}$ if Θ_I is finite), and the normal cone $N(\vartheta)$ is its orthogonal complement. Hence, letting Π_ϑ denote the orthogonal projection onto $N(\vartheta)$, it holds that $\|\Delta\|_\perp = \sup_{\vartheta\in\Theta_I} \|\Pi_\vartheta\Delta(\vartheta)\|$; in particular, if Θ_I is finite, $N(\vartheta) = \mathbb{R}^{d_\theta}$ and Π_ϑ is the identity.

Lemma 4.3 (Hausdorff distance between moved sets)**.** *Suppose that $\Theta_I \in \mathcal{K}_\Theta$ satisfies the following condition: there exists $r > 0$ such that every point of $\mathbb{R}^{d_\theta}$ within distance r of Θ_I has a unique nearest point in Θ_I. Then for any compact set $\mathcal{C}_0 \subset \mathcal{C}(\Theta_I)$, it holds that*

$$\lim_{n\to\infty} \sup_{g,g'\in\mathcal{C}_0} \left|\sqrt{n}\cdot d_H\left(\Theta_I \oplus_{1/\sqrt{n}} g,\ \Theta_I \oplus_{1/\sqrt{n}} g'\right) - \|g - g'\|_\perp\right| = 0.$$

Proof. See Appendix B.2. □

Lemma 4.3 shows that, to the first order, the distance between the two deviated sets depends on $g - g'$ only through the normal cones: at each $\vartheta \in \Theta_I$, only the part of $g(\vartheta) - g'(\vartheta)$ in the directions of $N(\vartheta)$ matters, while moving the points of Θ_I along Θ_I, i.e. in the directions of the tangent cone $T(\vartheta)$, does not change the set to the first order. For example, if Θ_I is a sphere and $g(\vartheta)$ is tangent to it at every ϑ, then $\|g\|_\perp = 0$ although $\Theta_I \oplus_1 g \neq \Theta_I$, while for a finite set $\Theta_I = \{\theta_1, \cdots, \theta_J\}$ it holds that $\|g - g'\|_\perp = \max_j \|g(\theta_j) - g'(\theta_j)\|$. Note that $\|\cdot\|_\perp$ is a seminorm on $\mathcal{C}(\Theta_I)$ with $\|\Delta\|_\perp \le \|\Delta\|_\infty$.

Positive reach, the condition in Lemma 4.3, is a standard regularity condition on sets (Federer, 1959). It holds for finite sets, for compact convex sets and for the identified sets considered in this paper, including those in $\mathcal{K}_1$, $\mathcal{K}_2$ and the full-dimensional sets in $\mathcal{K}_3$ (Lemma B.5 in Appendix B.2), and it excludes, e.g., sets with an inward corner. For sets with non-empty interior, $N(\vartheta) = \{0\}$ at every interior point ϑ of Θ_I, so that only the displacements of the boundary points matter. At a boundary point ϑ where the

[14]Such $\mathcal{G}$ exists; e.g. $\mathcal{G}(s) := \sum_k \langle s, e_k\rangle Z_k$, where $\{e_k\}_k$ is an orthonormal basis of $\{s \in L^2(P_0) : P_0 s = 0\}$, which is separable as $\mathcal{X}$ is a Borel subset of a Euclidean space, and $\{Z_k\}_k$ are independent $N(0,1)$ random variables.

boundary is smooth, $N(\vartheta)$ is the half-line spanned by the outward unit normal vector at ϑ, so that $\sup\{|u^t\Delta(\vartheta)| : u \in N(\vartheta), \|u\| \leq 1\}$ is the absolute value of the normal component of $\Delta(\vartheta)$: an outward and an inward displacement of the boundary count in the same way, as d_H is symmetric.

Theorem 4.4 (Asymptotic representation for set estimators)**.** *Suppose that the following conditions hold:*

(i). *Assumption 4.2 holds,*

(ii). *a set-valued statistic $T_n : \mathcal{X}^n \mapsto \mathcal{K}_\Theta$ satisfies that there exist measurable maps $\hat{g}_n : \mathcal{X}^n \mapsto \mathcal{C}(\Theta_I)$ such that, under P_0^n, $\sqrt{n}\cdot d_H(T_n, \Theta_I \oplus_{1/\sqrt{n}} \hat{g}_n) \to 0$ in probability and $\{\hat{g}_n\}_{n\in\mathbb{N}}$ is uniformly tight in $\mathcal{C}(\Theta_I)$.*

Then for any finite set $D \subset \Theta_I \times \mathbb{R}^{d_\theta}$ there exist a subsequence $\{n_k\}_{k\in\mathbb{N}}$, an independent random variable U uniformly distributed on $[0,1]$ and a randomized statistic $T = T(\mathcal{G}, U)$ with values in $\mathcal{C}(\Theta_I)$ such that, for every $(\theta, h) \in D$,

$$\mathcal{L}\left(\sqrt{n_k}\cdot d_H\left(T_{n_k}, \Theta_I(P_{\theta+h/\sqrt{n_k}})\right)\right) \rightsquigarrow \mathcal{L}\left(\|T - g_{\theta,h}\|_\perp\right) \quad \text{as } k \to \infty, \tag{18}$$

where the law on the left is taken under $P^{n_k}_{\theta+h/\sqrt{n_k}}$ and the law on the right under $\mathcal{G} \sim G_s$ with $s = \dot{\ell}_{\theta,h}h$.

Proof. See Appendix B.2. □

The implication of Theorem 4.4 is similar to the point identified case. For any set-valued estimator T_n satisfying the condition (ii), the law of the (scaled) Hausdorff distance $\sqrt{n}\cdot d_H(T_n, \Theta_I(P_{\theta+h/\sqrt{n}}))$ between T_n and the identified set under the locally deviated model is expressed by that of $\|T - g_{\theta,h}\|_\perp$ for a randomized statistic T in the limit experiment (16). Thus, the asymptotic risk of any such estimator T_n is bounded below by that of the statistic T in the limit experiment, which is the basis of the lower bound in Section 4.2. Note that the condition (ii) plays the role of the localization $\sqrt{n}(T_n - \theta_0)$ in the point identified case: $\hat{g}_n$ is the local displacement field of T_n, which does not depend on the local deviation (θ, h).

4.2. **Asymptotic lower bound.** I here derive the lower bound of the asymptotic risk arising from $\sqrt{n}\cdot d_H(T_n, \Theta_I(P_{\theta+h/\sqrt{n}}))$, i.e., the Hausdorff distance between T_n and the identified set under the locally deviated model. The classical results for the point identified case $\Theta_I = \{\theta_0\}$ have a long history (e.g. Hájek (1970), Hájek (1972)). As in this literature, the driving force here is the appropriate limit experiment obtained in Theorem 4.4. An intuition from the point identified case suggests that the limit

$$\mathcal{L}\left(\sqrt{n}\cdot d_H\left(T_n, \Theta_I(P_{\theta+h/\sqrt{n}})\right)\right) \overset{a}{\sim} \mathcal{L}\left(\|T(\mathcal{G}, U) - g_{\theta,h}\|_\perp\right), \quad \mathcal{G} \sim G_s, \quad s = \dot{\ell}_{\theta,h}h,$$

provides the lower bound of the corresponding asymptotic risk. The departure from the point identified case is that the target in the limit is the displacement field $g_{\theta,h}$ on Θ_I, and that the distance $\|\cdot\|_\perp$ takes the supremum of the normal displacements over Θ_I. This intuition plays a key role when deriving the asymptotic lower bound of the risk.

I proceed with my analysis from the local minimax perspective, i.e., the central question here is "how accurately can one estimate the identified set $\Theta_I(P_{\theta+h/\sqrt{n}})$ when the data are generated by the locally deviated model $P^n_{\theta+h/\sqrt{n}}$?" and I focus on the worst case of this accuracy over the local deviations $(\theta, h) \in \Theta_I \times \mathbb{R}^{d_\theta}$. Since the limit experiment in Theorem 4.4 depends on (θ, h) only through the score $\dot{\ell}_{\theta,h}h$, the worst case is effectively taken over the parameter of the limit experiment; when this parameter is finite dimensional and the target field is linear in it, the limit experiment is a Gaussian shift and the minimax problem can be solved explicitly. This intuition is formalized in Theorem 4.6.

Assumption 4.5 (Linear local deviation)**.** There exist $d_a \in \mathbb{N}$, linearly independent $s_1, \cdots, s_{d_a} \in L^2(P_0)$ and a linear map $\Upsilon : \mathbb{R}^{d_a} \mapsto \mathcal{C}(\Theta_I)$ such that $\{\dot{\ell}_{\theta,h}h : \theta \in \Theta_I,\ h \in \mathbb{R}^{d_\theta}\} = \{\sum_{j=1}^{d_a} a_j s_j : a \in \mathbb{R}^{d_a}\}$ and $g_{\theta,h} = \Upsilon a$ whenever $\dot{\ell}_{\theta,h}h = \sum_{j=1}^{d_a} a_j s_j$.

Assumption 4.5 concerns how the local deviations $(\theta, h) \in \Theta_I \times \mathbb{R}^{d_\theta}$ are seen through the data. The first condition in Assumption 4.5 requires that the scores of all the local deviations form a finite dimensional linear space. Since $s_1, \cdots, s_{d_a}$ are linearly independent, each score is written as $\dot{\ell}_{\theta,h}h = \sum_{j=1}^{d_a} a_j s_j$ with a unique $a \in \mathbb{R}^{d_a}$, so that a can be seen as the local parameter revealed by the data, with the Fisher information $I_a := (P_0 s_j s_k)_{j,k=1}^{d_a}$. The second condition requires that the target field $g_{\theta,h}$ depends on (θ, h) only through a, and linearly. The first part of this requirement is essential: since the limit experiment in Theorem 4.4 depends on (θ, h) only through the score, two local deviations with the same score cannot be distinguished even in the limit, and no estimator can be close to both identified sets if their target fields differ. Under Assumption 4.5, the limit experiment reduces to the Gaussian shift experiment $\{N(a, I_a^{-1}) : a \in \mathbb{R}^{d_a}\}$ with the linear target Υa, which allows one to solve the minimax problem explicitly.

Assumption 4.5 holds in standard cases. In the point identified case $\Theta_I = \{\theta_0\}$ with $I_{\theta_0} > 0$, it holds with $d_a = d_\theta$, $s_1, \cdots, s_{d_a}$ the components of $\dot{\ell}_{\theta_0}$ and $\Upsilon a = a$, so that $I_a = I_{\theta_0}$.[15]

Theorem 4.6 (Local asymptotic minimax bound for Θ_I)**.** *Suppose that Assumptions 4.2 and 4.5 hold, and let $I_a := (P_0 s_j s_k)_{j,k=1}^{d_a}$. Then for any set-valued estimators $T_n : \mathcal{X}^n \mapsto \mathcal{K}_\Theta$ satisfying the condition (ii) in Theorem 4.4 and for any non-decreasing and lower semicontinuous loss function $l : [0, \infty) \mapsto [0, \infty)$, it holds that*

$$\sup_{D \in \mathcal{D}} \liminf_{n \to \infty} \sup_{(\theta,h) \in D} \mathbb{E}_{P^n_{\theta+h/\sqrt{n}}} \left[l \circ \left(\sqrt{n} \cdot d_H \left(T_n, \Theta_I(P_{\theta+h/\sqrt{n}}) \right) \right) \right] \geq \mathbb{E} \left[l \circ \| \Upsilon Z \|_\perp \right],$$

where $Z \sim N(0, I_a^{-1})$ and $\mathcal{D} := \{D \subset \Theta_I \times \mathbb{R}^{d_\theta} : 0 < |D| < \infty\}$ is a directed set with the partial order $\subset$.

[15] In the example of footnote 9, the score is $\dot{\ell}_{\theta_1,h}h = (\theta_1 - \theta_2)hX$ and $\dot{\ell}_{\theta_2,h}h = -(\theta_1 - \theta_2)hX$, so that Assumption 4.5 holds with $d_a = 1$, $s_1 = X$, $I_a = 1$ and $(\Upsilon a)(\theta_1) = a/(\theta_1 - \theta_2)$, $(\Upsilon a)(\theta_2) = -a/(\theta_1 - \theta_2)$.

Proof. See Appendix B.3. □

Theorem 4.6 shows that the Gaussian shift experiment under Assumption 4.5 provides a generic lower bound of the asymptotic risk of $\sqrt{n} \cdot d_H(T_n, \Theta_I(P_{\theta+h/\sqrt{n}}))$ in a minimax sense. Roughly speaking, the lower bound is governed by the information I_a and the linear map Υ; smaller information I_a makes it harder to estimate the local parameter a, and Υ translates the estimation error Z of a into the displacement $(\Upsilon Z)(\vartheta)$ of each point $\vartheta \in \Theta_I$, of which only the normal component $\Pi_\vartheta(\Upsilon Z)(\vartheta)$ matters. Note that the supremum over $\vartheta \in \Theta_I$ in $\|\Upsilon Z\|_\perp$ is taken inside the expectation: since the displacements of all the points of Θ_I are driven by the same Z, the bound accounts for them jointly, and it is at least as large as $\sup_{\vartheta \in \Theta_I} \mathbb{E}[l(\|\Pi_\vartheta(\Upsilon Z)(\vartheta)\|)]$, the bound obtained by looking at each point separately; the two coincide when all the points move by the same normal amount, as in the example of footnote 9.

Remark 4.7. Theorem 4.6 nests the classical result for the point identified case. When $\Theta_I = \{\theta_0\}$ with $I_{\theta_0} > 0$, the locally deviated model in (12) reduces to $\{P^n_{\theta_0+h/\sqrt{n}} : h \in \mathbb{R}^{d_\theta}\}$ with $\Theta_I(P_{\theta_0+h/\sqrt{n}}) = \{\theta_0 + h/\sqrt{n}\}$, and Assumption 4.5 holds with $\Upsilon a = a$ and $I_a = I_{\theta_0}$. For estimators that are singletons, $T_n = \{\hat{\theta}_n\}$, the Hausdorff distance reduces to the Euclidean distance, $d_H(T_n, \Theta_I(P_{\theta_0+h/\sqrt{n}})) = \|\hat{\theta}_n - \theta_0 - h/\sqrt{n}\|$, and the lower bound in Theorem 4.6 reduces to $\mathbb{E}[l \circ \|Z\|]$ with $Z \sim N(0, I_{\theta_0}^{-1})$, which corresponds to the local minimax result for the point estimation of θ_0 (e.g. Hájek (1972)). Note that Theorem 4.6 is stated for the estimators satisfying the condition (ii) in Theorem 4.4, i.e. $\sqrt{n}(T_n - \theta_0)$ being uniformly tight in the point identified case, whereas the classical result holds for any estimator.

Remark 4.8. My lower bound is based on the local asymptotic minimax criterion. In point identified cases, there is another argument to derive the lower bound via convolution (e.g. Hájek (1970)). Compared with the minimax criteria, this approach restricts the scope of estimators to those satisfying "local regularity" conditions. As pointed out in Hirano and Porter (2012), however, locally regular estimators need not exist when the parameter of interest is not differentiable, so that this requirement can be restrictive in the set identified context; the minimax criterion does not need such a requirement, and Theorem 4.6 applies to any estimator satisfying the condition (ii) in Theorem 4.4.

4.3. **General moment conditions.** I now consider the semiparametric case, where no likelihood function fully describes the model but moment restrictions are available, based on which the finite-dimensional parameters of interest are characterized. To derive the asymptotic efficiency bound in this case, I extend the analysis of the parametric model in Section 4.2 to a semiparametric one. For a point identified parameter defined by moment restrictions, the efficiency bound is due to Chamberlain (1987). I derive its counterpart for the identified set, using parametric submodels through the true distribution as in semiparametric efficiency theory (e.g. van der Vaart (1998, Chapter 25)).

The idea is as follows. Given a moment restriction $\Phi = q \circ (\mathbb{E}_P m)$ and the corresponding identified set defined as in (4), I consider finite-dimensional parametric submodels through the true distribution P_0, along which the distribution of the data moves in finitely many directions. Along such a submodel, the identified set moves by a field that is linear in the direction (Lemma 4.10), as required in Section 4.2, so that the results there give a lower bound for each submodel (Proposition 4.11). Letting the submodels span all the directions in which P_0 can move then gives the lower bound over a neighborhood of P_0, expressed through the moment restrictions alone (Theorem 4.12). Unlike in the point identified case, each step requires the treatment of the identified set as a whole, through the geometry in Lemma 4.3.

I first state the requirements on the moment restriction $\Phi = q \circ (\mathbb{E}_P m)$, based on which Θ_I is defined, and on the true distribution P_0. In this section I explicitly write $\Theta_I(\Phi, P)$ to make the dependence of Θ_I on Φ and P explicit. The regularity conditions on (Φ, P_0) are formally summarized in Assumption 4.9 as follows.

Assumption 4.9 (Regularity conditions on (Φ, P_0))**.** For $P \in \mathcal{P}$ and $\theta \in \Theta_I(\Phi, P)$, let $\mathcal{A}(\theta, P) := \{j \in \mathcal{J}_2 : \Phi_j(\theta, P) = 0\}$ denote the set of binding inequality restrictions in (4). The pair (Φ, P_0) satisfies the following conditions;

(i). $\emptyset \neq \Theta_I(\Phi, P_0) \subset \Theta^o$, and $\Theta_I(\Phi, P_0)$ satisfies the condition in Lemma 4.3,

(ii). $P_0\|m(\theta, \cdot)\| < \infty$ for $\theta \in \Theta$, $\theta \mapsto P_0 m(\theta, \cdot)$ is continuous on Θ and C^1 on a neighborhood of $\Theta_I(\Phi, P_0)$, and there exists a set $\mathcal{S}$ of bounded measurable functions $s : \mathcal{X} \mapsto \mathbb{R}$ with $P_0 s = 0$, whose linear span is dense in $\{s \in L^2(P_0) : P_0 s = 0\}$, such that for each $s \in \mathcal{S}$ the map $\theta \mapsto P_0[m(\theta, \cdot)s]$ is continuous on Θ and C^1 on a neighborhood of $\Theta_I(\Phi, P_0)$,

(iii). for each $\theta \in \Theta_I(\Phi, P_0)$, the gradients $\nabla_\theta \Phi_j(\theta, P_0)$, $j \in \mathcal{J}_1 \cup \mathcal{A}(\theta, P_0)$, are linearly independent,

(iv). q is continuous, and C^1 on a neighborhood of $\{P_0 m(\theta, \cdot) : \theta \in \Theta_I(\Phi, P_0)\}$,

(v). for each component m_i of m, $i = 1, \cdots, d_m$, the class $\mathcal{M}_i := \{m_i(\theta, \cdot)\}_{\theta \in \Theta_I(\Phi, P_0)}$ satisfies $\int_0^1 \sqrt{\log N_{[]}(\varepsilon, \mathcal{M}_i, L^2(P_0))}\, d\varepsilon < \infty$, where $N_{[]}(\varepsilon, \mathcal{M}_i, L^2(P_0))$ is the minimal number of brackets $\{f : f_L \leq f \leq f_U\}$ with $f_L, f_U \in L^2(P_0)$ and $P_0(f_U - f_L)^2 \leq \varepsilon^2$ needed to cover $\mathcal{M}_i$, and $\theta \mapsto m(\theta, \cdot)$ is continuous in $L^2(P_0)$ on $\Theta_I(\Phi, P_0)$.

Assumption 4.9 is imposed at the true distribution P_0 only. Its condition (i) requires that the identified set satisfies the condition in Lemma 4.3 (positive reach); no condition on other distributions is needed, since the lower bound below is derived along smooth paths through P_0, along which the behavior of the identified set follows from (i)–(iv) (Lemma 4.10 below). Assumption 4.9 (ii) places the smoothness on the moment function; together with (iv), it implies that $\Phi(\cdot, P_0)$ is C^1 on a neighborhood of $\Theta_I(\Phi, P_0)$. The second part of (ii) requires the same smoothness along the directions $s \in \mathcal{S}$ in which the distribution of the data moves below[16]. Assumption 4.9 (iii) is the

[16]The second part of (ii) holds with $\mathcal{S}$ the set of all bounded functions with mean zero when $\theta \mapsto m(\theta, x)$ is C^1 for each x with an integrable envelope for $\nabla_\theta m$, as under Assumption 3.4 in Kaido

Linear Independence Constraint Qualification (LICQ) for the binding restrictions[17]. It covers moment inequality models, whose identified set typically has non-empty interior and may have corners, as well as sets defined by exactly identifying moment equalities, including just identified point identification. It excludes over-identifying moment equalities[18]; see Remark 4.14 for the over-identified point identified case. Under (ii)–(iv), the identified set near each of its points is determined by the binding restrictions alone, and it moves linearly when P moves along a path of distributions (Lemma 4.10 below), which is what Assumption 4.5 requires in the present setting. Assumption 4.9 (ii) and (iv) make $\Phi(\theta, \cdot)$ differentiable along such paths, with a derivative that is linear in the score of the path. Assumption 4.9 (v) is the standard bracketing entropy condition, which implies that $\{m(\theta, \cdot)\}_{\theta \in \Theta_I(\Phi, P_0)}$ is P_0-Donsker (van der Vaart (1998, Theorem 19.5); cf. Assumption 5.1 (ii)).

I now consider local deviations of the true distribution $P_0 \in \mathcal{P}$ along smooth parametric paths. Let $s_1, \cdots, s_{d_a} : \mathcal{X} \mapsto \mathbb{R}$ be bounded measurable functions with $P_0 s_j = 0$ that are linearly independent in $L^2(P_0)$. For $a \in \mathbb{R}^{d_a}$ and $t \geq 0$ with $t \cdot \sup_{x \in \mathcal{X}} |\sum_{j=1}^{d_a} a_j s_j(x)| < 1$, let $P_{t,a}$ be the probability measure on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ given by

$$dP_{t,a} := \left(1 + t \sum\nolimits_{j=1}^{d_a} a_j s_j\right) dP_0.$$

The path $t \mapsto P_{t,a}$ passes through P_0 at $t = 0$ and has the score $\sum_{j=1}^{d_a} a_j s_j$; see section 25.3 in van der Vaart (1998). The paths and their scores play the roles of $t \mapsto P_{\theta + th}$ and $\dot{\ell}_{\theta,h} h$ in Section 4.1.

Along these paths, the identified set (4) is $\Theta_I(\Phi, P_{t,a})$ with

$$\Phi(\theta, P_{t,a}) = q\left(P_0 m(\theta, \cdot) + t \sum\nolimits_{j=1}^{d_a} a_j P_0 \left[m(\theta, \cdot) s_j\right]\right).$$

By Assumption 4.9 (ii) and (iv), for θ near $\Theta_I(\Phi, P_0)$ the map $t \mapsto \Phi(\theta, P_{t,a})$ is differentiable at $t = 0$, and its derivative is linear in a. The following lemma shows that along these paths the identified set moves linearly in a; it is the counterpart of Assumption 4.2 (iii) and Assumption 4.5 in the present setting.

Lemma 4.10 (Local deviation of the identified set along smooth paths)**.** *Let $\Theta_I := \Theta_I(\Phi, P_0)$, and suppose that Assumption 4.9 holds and that $s_1, \cdots, s_{d_a} \in \mathcal{S}$. Then there exists a linear map $\Upsilon : \mathbb{R}^{d_a} \mapsto \mathcal{C}(\Theta_I)$ such that, for every $a \in \mathbb{R}^{d_a}$,*

(a). $\nabla_\theta \Phi_j(\vartheta, P_0)^t (\Upsilon a)(\vartheta) = -\frac{d}{dt} \Phi_j(\vartheta, P_{t,a})\big|_{t=0}$ *for every $\vartheta \in \Theta_I$ and $j \in \mathcal{J}_1 \cup \mathcal{A}(\vartheta, P_0)$,*

(b). $\lim_{n \to \infty} \sqrt{n} \cdot d_H\left(\Theta_I(\Phi, P_{1/\sqrt{n}, a}),\ \Theta_I \oplus_{1/\sqrt{n}} \Upsilon a\right) = 0.$

Proof. See Appendix B.4. □

and Santos (2014), and with $\mathcal{S}$ the set of continuous bounded functions with mean zero for indicator moment functions such as $m(\theta, x) = 1\{x \leq \theta\}$ when P_0 has a continuous density near $\Theta_I(\Phi, P_0)$.

[17] See Kaido et al. (2022) for its role in partially identified models, and Assumption 3.6 (iv) in Kaido and Santos (2014) for the same condition in convex moment inequality models.

[18] The gradients of more than d_θ equalities cannot be linearly independent.

Lemma 4.10 (b) is the counterpart of (14) with $g = \Upsilon a$, and the local deviation Υa is linear in the direction a, as required in Assumption 4.5. The map Υ is not unique: (a) fixes only the components of $(\Upsilon a)(\vartheta)$ along the gradients of the binding restrictions. These are, however, the only components that matter in Lemma 4.3: under Assumption 4.9 (i)–(iv), every $u \in N(\vartheta)$ is a linear combination of $\nabla_\theta \Phi_j(\vartheta, P_0)$, $j \in \mathcal{J}_1 \cup \mathcal{A}(\vartheta, P_0)$, with non-negative coefficients on $j \in \mathcal{A}(\vartheta, P_0)$ (Rockafellar and Wets (1998, Theorem 6.14)), so that $u^t(\Upsilon a)(\vartheta)$ is determined through (a) by the rates at which the binding restrictions change along the path.

Lemma 4.10 provides, along the paths $t \mapsto P_{t,a}$, the local deviation of the identified set required in Theorems 4.4 and 4.6. This leads to the following lower bound for $\Theta_I(\Phi, P_0)$.

Proposition 4.11 (Local asymptotic minimax bound for Θ_I along smooth paths)**.** *Let $\Theta_I := \Theta_I(\Phi, P_0)$, suppose that the assumptions of Lemma 4.10 hold, and let Υ be as in Lemma 4.10. Then for any set-valued estimators $T_n : \mathcal{X}^n \mapsto \mathcal{K}_\Theta$ satisfying the condition (ii) in Theorem 4.4 and for any non-decreasing and lower semicontinuous loss function $l : [0, \infty) \mapsto [0, \infty)$, it holds that*

$$\sup_{D \in \mathcal{D}} \liminf_{n \to \infty} \sup_{a \in D} \mathbb{E}_{P^n_{1/\sqrt{n},a}} \left[l \circ \left(\sqrt{n} \cdot d_H \left(T_n, \Theta_I(\Phi, P_{1/\sqrt{n},a}) \right) \right) \right] \geq \mathbb{E} \left[l \circ \| \Upsilon Z \|_\perp \right],$$

where $Z \sim N(0, I_a^{-1})$ with $I_a := (P_0 s_j s_k)_{j,k=1}^{d_a}$, and $\mathcal{D} := \{D \subset \mathbb{R}^{d_a} : 0 < |D| < \infty\}$ is a directed set with the partial order $\subset$.

Proof. See Appendix B.4. □

Proposition 4.11 is the counterpart of Theorem 4.6 for the paths $t \mapsto P_{t,a}$: the local parameter $a \in \mathbb{R}^{d_a}$ is the direction in which the distribution of the data moves, and I_a is the information of the scores $s_1, \cdots, s_{d_a}$ as in Assumption 4.5. The bound does not depend on the choice of Υ in Lemma 4.10, since $\| \Upsilon Z \|_\perp$ depends on ΥZ only through the components fixed by Lemma 4.10 (a).

Proposition 4.11 provides a lower bound for each finite set of scores $s_1, \cdots, s_{d_a}$, that is, for each finite-dimensional parametric submodel through P_0. Since the worst-case risk over a neighborhood of P_0 is at least each of these bounds, I next let the scores span all the directions in which P_0 can move. To this end, equip the set of all probability measures on $\mathcal{X}$ with the topology τ of Chamberlain (1987), induced by the basic neighborhoods of P_0 defined as

$$\mathcal{N}_{P_0} := \left\{ P : \left| \int_{\mathcal{X}} f_j(x) \left(dP(x) - dP_0(x) \right) \right| < \epsilon_j, \ j = 1, \cdots, k \right\},$$

where P ranges over the probability measures on $\mathcal{X}$, $\epsilon_j > 0$, $k \in \mathbb{N}$, and $f_j : \mathcal{X} \mapsto \mathbb{R}$ are measurable functions with $f_j \in L^1(P_0)$[19]. The bound involves the Gaussian process $\mathcal{G}$ of Section 4.1, applied componentwise to vector-valued functions. Under P_0, $\mathcal{G}(s)$ is the limit in distribution of $\mathbb{G}_n s$ in (15), so that, e.g., $\mathcal{G}(m(\vartheta, \cdot) - P_0 m(\vartheta, \cdot))$ is the limit in distribution of $\sqrt{n}(\mathbb{P}_n - P_0) m(\vartheta, \cdot)$. The formal result is stated as follows.

[19]This topology is different from the weak topology, which is defined via continuous and bounded functions $f : \mathcal{X} \mapsto \mathbb{R}$.

Theorem 4.12 (Local asymptotic minimax bound for Θ_I: a general case)**.** *Let $\Theta_I := \Theta_I(\Phi, P_0)$, suppose that Assumption 4.9 holds, and let V be a random element of $\mathcal{C}(\Theta_I)$ such that, for every $\vartheta \in \Theta_I$ and $j \in \mathcal{J}_1 \cup \mathcal{A}(\vartheta, P_0)$,*

$$\nabla_\theta \Phi_j(\vartheta, P_0)^t V(\vartheta) = -\nabla q_j\left(P_0 m(\vartheta, \cdot)\right) \mathcal{G}\left(m(\vartheta, \cdot) - P_0 m(\vartheta, \cdot)\right). \tag{19}$$

Then for any set-valued estimators $T_n : \mathcal{X}^n \mapsto \mathcal{K}_\Theta$ satisfying the condition (ii) in Theorem 4.4, for any non-decreasing and lower semicontinuous loss function $l : [0, \infty) \mapsto [0, \infty)$ and for any basic neighborhood $\mathcal{N}_{P_0}$ of P_0, it holds that

$$\liminf_{n\to\infty} \sup_{P \in \mathcal{N}_{P_0} : \Theta_I(\Phi, P) \in \mathcal{K}_\Theta} \mathbb{E}_{P^n}\left[l \circ \left(\sqrt{n} \cdot d_H\left(T_n, \Theta_I(\Phi, P)\right)\right)\right] \geq \mathbb{E}\left[l \circ \|V\|_\perp\right].$$

Proof. See Appendix B.4. □

Theorem 4.12 extends Theorem 4.6 and Proposition 4.11 to the case where the distribution of the data is restricted only through Assumption 4.9. The bound has the same form as before, with ΥZ replaced by the field V, whose components along the gradients of the binding restrictions are given by the Gaussian limit of the influence functions of these restrictions in (19). Such a field V exists, and $\mathbb{E}[l \circ \|V\|_\perp]$ does not depend on its choice, for the same reason as for Υ after Lemma 4.10. The right-hand side of (19) is minus the limit in distribution of $\sqrt{n}\{\Phi_j(\vartheta, \mathbb{P}_n) - \Phi_j(\vartheta, P_0)\}$ under P_0, that is, of the normalized estimation error of the j-th restriction at ϑ. Hence, at a point ϑ where a single restriction j binds, the component $u^t V(\vartheta)$ of $V(\vartheta)$ along the unit normal $u := \nabla_\theta \Phi_j(\vartheta, P_0)/\|\nabla_\theta \Phi_j(\vartheta, P_0)\|$ equals the right-hand side of (19) divided by $\|\nabla_\theta \Phi_j(\vartheta, P_0)\|$: the boundary moves along its normal by the noise in the binding restriction divided by its slope. Example 4.13 below illustrates the bound.

The proof applies Proposition 4.11 to finite sets of scores and lets the scores span all directions; the bound therefore holds already when the worst case is taken over the smooth paths of Proposition 4.11 only. If the functions $\nabla q_j(P_0 m(\vartheta, \cdot))(m(\vartheta, \cdot) - P_0 m(\vartheta, \cdot))$, $\vartheta \in \Theta_I$, $j \in \mathcal{J}_1 \cup \mathcal{A}(\vartheta, P_0)$, lie in the span of finitely many $s_1, \cdots, s_{d_a} \in \mathcal{S}$, as when the data enter the moment function only through finitely many fixed functions, e.g. $m(\theta, y) = g(y) - c(\theta)$, then Proposition 4.11 with these scores already gives the bound in Theorem 4.12, since the projection in its proof leaves the noise in the binding restrictions unchanged; letting the scores span all directions is needed when this noise varies with ϑ in infinitely many directions, as for moment functions that are nonlinear in θ on a boundary with infinitely many points. The bound is valid whether Θ_I is a set or, under exactly identifying moment equalities, a point; see Remark 4.14.

Example 4.13 (Two means)**.** Let $Y = (Y_1, Y_2)$ with $Y_1 \sim N(\mu_1, 1)$ and $Y_2 \sim N(\mu_2, 1)$ independent, where $\mu_1 < \mu_2$. Take $m(\theta, y) = (y_1 - \theta, \theta - y_2)^t$, q the identity and $\mathcal{J}_1 = \emptyset$, so that

$$\Theta_I(\Phi, P) = \{\theta \in \Theta : PY_1 \leq \theta \leq PY_2\}$$

and $\Theta_I = [\mu_1, \mu_2]$ under P_0. With Θ a compact interval containing $[\mu_1, \mu_2]$ in its interior, Assumption 4.9 holds.

Only the two endpoints matter for the bound. At μ_1, only the first restriction binds, with gradient -1, so that $N(\mu_1) = (-\infty, 0]$. At μ_2, only the second restriction binds,

with gradient 1, so that $N(\mu_2) = [0, \infty)$. At interior points, no restriction binds and $N(\vartheta) = \{0\}$. Let Z_1 and Z_2 be the values of $\mathcal{G}$ at $y \mapsto y_1 - \mu_1$ and $y \mapsto y_2 - \mu_2$. They are the limits in distribution of $\sqrt{n}(\bar{Y}_1 - \mu_1)$ and $\sqrt{n}(\bar{Y}_2 - \mu_2)$, and they are independent $N(0, 1)$. At the endpoints, (19) reads

$$V(\mu_1) = Z_1, \quad V(\mu_2) = Z_2.$$

That is, each endpoint moves by the noise in its own mean, as both restrictions have slope 1 in absolute value. Hence $\|V\|_\perp = \max\{|Z_1|, |Z_2|\}$ for any V satisfying (19), and the bound in Theorem 4.12 is

$$\mathbb{E}\left[l\left(\max\{|Z_1|, |Z_2|\}\right)\right],$$

which equals $2/\sqrt{\pi}$ for $l(z) = z$.

The estimator $T_n = [\bar{Y}_1, \bar{Y}_2]$ has limiting risk under P_0 equal to this bound. Indeed, $\bar{Y}_1 \leq \bar{Y}_2$ with probability approaching one, and then

$$\sqrt{n} \cdot d_H(T_n, \Theta_I) = \max\left\{\sqrt{n}|\bar{Y}_1 - \mu_1|,\ \sqrt{n}|\bar{Y}_2 - \mu_2|\right\}.$$

If Y_1 and Y_2 are correlated, the theorem applies unchanged; only the joint law of (Z_1, Z_2) changes, through the covariance $P_0[(Y_1 - \mu_1)(Y_2 - \mu_2)]$.

Remark 4.14. Theorem 4.12 contains the efficiency bound of Chamberlain (1987) for a just identified, point identified parameter. Let $d_m = d_q = d_\theta$, let q be the identity and $\mathcal{J}_1 = \{1, \cdots, d_\theta\}$, so that $\Theta_I(\Phi, P) = \{\theta \in \Theta : Pm(\theta, \cdot) = 0\}$, and suppose that $\Theta_I(\Phi, P_0) = \{\theta_0\}$ and that Assumption 4.9 holds. Then Assumption 4.9 (iii) says that $G := \nabla_\theta P_0 m(\theta_0, \cdot)$ is non-singular, $N(\theta_0) = \mathbb{R}^{d_\theta}$, and (19) gives $V(\theta_0) = -G^{-1}\mathcal{G}(m(\theta_0, \cdot)) \sim N(0, G^{-1}\Omega G^{-t})$ with $\Omega := P_0 m(\theta_0, \cdot) m(\theta_0, \cdot)^t$, so that $\|V\|_\perp = \|V(\theta_0)\|$. For estimators $T_n = \{\hat{\theta}_n\}$, $d_H(T_n, \{\theta_0\}) = \|\hat{\theta}_n - \theta_0\|$, and the bound in Theorem 4.12 is $\mathbb{E}[l(\|Z\|)]$ with $Z \sim N(0, G^{-1}\Omega G^{-t})$. When Ω is non-singular, $G^{-1}\Omega G^{-t} = (G^t \Omega^{-1} G)^{-1}$ is the efficiency bound in Chamberlain (1987), which is attained by the GMM estimator. With more moment equalities than parameters, the gradients of the equalities cannot be linearly independent, so that Assumption 4.9 (iii) fails. The extension to cover this over-identified singleton case is left for future research.

5. Limiting Distribution of Set Estimator from Criterion Function

I now consider a set estimator T_n of Θ_I. As in Section 4, I remain agnostic about the characteristics of identified sets beyond Assumption 3.3, for the purpose of adaptiveness. The analysis in Section 4, in particular in Section 4.3, indicates that T_n could be defined as the extremum estimator of a criterion function of GMM-type $Q : \Theta \mapsto \mathbb{R}_+$ where Θ_I is defined as the zero-level set of Q. In fact, this type of criterion function has been studied in the literature (e.g. Chernozhukov et al. (2007) and Romano and Shaikh (2010)). The main result of this section derives the limiting distribution of the extremum-type set estimators from this criterion function in the Hausdorff metric d_H.

In Section 5.1, I first set up the criterion function, define the set estimator, and show its consistency and rate of convergence. Based on this result, in Section 5.2 I derive its limiting distribution in terms of the Hausdorff metric d_H.

5.1. **Setting and basic properties of set estimator $\widehat{\Theta}_I$.** Throughout this section, I work with the expression (4) of $\Theta_I = \Theta_I(\Phi, P)$, with the moment equality restrictions $\Phi_{\mathcal{J}_1}$ and the moment inequality restrictions $\Phi_{\mathcal{J}_2}$ of Section 3.1; P denotes the true distribution, written P_0 in Section 4.3.

From the model and the identified set expressed above, define the criterion function $Q : \Theta \mapsto \mathbb{R}_+$ such that

$$Q(\theta) = \left\| \Phi_v^t(\theta, P) \cdot W(\theta)^{1/2} \right\|, \quad \theta \in \Theta, \tag{20}$$

where $W(\theta) \in \mathbb{R}^{d_q \times d_q}$ is a positive definite matrix for each $\theta \in \Theta$, continuous in θ, and

$$\Phi_v^t(\theta, P) = \left(\Phi_{\mathcal{J}_1}^t(\theta, P),\ [\Phi_{\mathcal{J}_2}^t(\theta, P)]_+ \right) = \left(\left(\Phi_j^t(\theta, P) \right)_{j \in \mathcal{J}_1},\ \left([\Phi_j^t(\theta, P)]_+ \right)_{j \in \mathcal{J}_2} \right),$$

with $[\Phi_j(\theta, P)]_+ := (\Phi_j(\theta, P) \vee 0)$ for any $j \in \mathcal{J}_2$. Clearly, $\Theta_I = \{\theta \in \Theta : \Phi_{\mathcal{J}_1}(\theta, P) = 0, \Phi_{\mathcal{J}_2}(\theta, P) \leq 0\} = \{\theta \in \Theta : Q(\theta) \leq 0\}$.

The set estimator, denoted by $\widehat{\Theta}_I(\tau)$ with the threshold level $\tau \in \mathbb{R}_+$ which could be data-dependent, is a lower level set of the sample analogue of (20), i.e.,

$$\widehat{\Theta}_I(\tau) = \{\theta \in \Theta : a_n \cdot Q_n(\theta) \leq \tau\}, \tag{21}$$

where $\{a_n\}_{n \in \mathbb{N}}$ is a normalizing sequence and

$$Q_n(\theta) = \left\| \Phi_v^t(\theta, \mathbb{P}_n) \cdot W_n(\theta)^{1/2} \right\|, \quad \theta \in \Theta, \tag{22}$$

where $W_n(\theta)$ is the sample weighting matrix and

$$\Phi_v^t(\theta, \mathbb{P}_n) = \left(\Phi_{\mathcal{J}_1}^t(\theta, \mathbb{P}_n),\ [\Phi_{\mathcal{J}_2}^t(\theta, \mathbb{P}_n)]_+ \right) = \left(\left(\Phi_j^t(\theta, \mathbb{P}_n) \right)_{j \in \mathcal{J}_1},\ \left([\Phi_j^t(\theta, \mathbb{P}_n)]_+ \right)_{j \in \mathcal{J}_2} \right),$$

with $\mathbb{P}_n$ an empirical measure on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. Regarding the choice of τ, a key quantity is

$$c_n := \inf_{\theta' \in \Theta} Q_n(\theta'). \tag{23}$$

The setting above is standard. The vector $\Phi_v(\theta, P)$ represents deviation from constraint, which treats $\Phi_{\mathcal{J}_1}(\theta, P)$ and $\Phi_{\mathcal{J}_2}(\theta, P)$ differently. The quantity c_n is important in terms of self-adjustment of the threshold level which determines the estimator $\widehat{\Theta}_I$; in moment equality with over identified point it holds that $c_n > 0$. Note that this type of classic criterion function potentially appears to kill the information of identified set in a sense that differentiability does not hold and Hadamard directional derivatives of $\theta \in \Theta_I$ vanish (with respect to some direction). Note that, even though this criterion function is not fully differentiable or some direction of Hadamard directional derivative vanishes, the key in deriving limiting distribution is differentiability of Φ; this point is discussed in detail in Section 5.2.

The consistency and rate of convergence of $\widehat{\Theta}_I$ can be derived via the approach in Chernozhukov et al. (2007). The next assumptions are used to analyze the consistency, the convergence rate and the limiting distribution of $\widehat{\Theta}_I$.

Assumption 5.1. (i). The functions $Q : \Theta \mapsto \mathbb{R}_+$ and $Q_n : \Theta \times \mathcal{X} \mapsto \mathbb{R}_+$ are Borel measurable. (ii). The class of moment functions $\{m(\theta, \cdot)\}_{\theta \in \Theta}$ has a P-Donsker property; uniformly over $\theta \in \Theta$ it holds that

$$\mathbb{G}_n[m(\theta, \cdot)] := \sqrt{n} \left(\mathbb{P}_n m(\theta, \cdot) - P m(\theta, \cdot) \right) \rightsquigarrow \xi(\theta), \quad \text{as } n \to \infty, \tag{24}$$

where $\xi(\theta) := \mathcal{G}(m(\theta, \cdot) - Pm(\theta, \cdot))$, with $\mathcal{G}$ the Gaussian process of Section 4.1, taken with $P_0 = P$ and applied componentwise. (iii). $W_n(\theta) = W(\theta) + o_P(1)$ uniformly over $\theta \in \Theta$, where $W(\theta)$ is positive definite and continuous for any $\theta \in \Theta$.

Assumption 5.2 (The "degeneracy" condition in Chernozhukov et al. (2007).)**.** (i). There is a sequence of subsets $\{\Theta_n\}_{n \in \mathbb{N}} \subset \Theta$, which could be data dependent, such that $Q_n(\theta) - \inf_{\theta' \in \Theta} Q_n(\theta') = 0$ for each $\theta \in \Theta_n$ and for each $n \in \mathbb{N}$, and that $d_H(\Theta_n, \Theta_I) \leq \epsilon_n$ for some $\epsilon_n = o_P(1)$. (ii). The sequence ϵ_n is of stochastic order $a_n^{-1/\gamma}$ for some $\gamma > 0$.

Assumption 5.1 is a set of standard regularity conditions; a sufficient condition for its part (ii) is the bracketing entropy condition in Assumption 4.9 (v) with $\Theta_I(\Phi, P_0)$ replaced by Θ (van der Vaart (1998, Theorem 19.5)). Assumption 5.2 is satisfied in point identified models and many moment inequality models as shown in Chernozhukov et al. (2007)[20]. There are two quantities a_n and γ. First, recall that a_n appears as the scaling factor of Q_n in (22); in the present context, the order $a_n = \sqrt{n}$ is appropriate given the P-Donsker condition in Assumption 5.1 and the construction of Q in (20). Second, the quantity $\gamma > 0$ in Assumption 5.2 (ii) is related to the one in "Polynomial minorant" condition, Condition C.2 of Chernozhukov et al. (2007). Assumption 4.9 implies, through the LICQ in its condition (iii), the partial identification conditions (4.1) and (4.5) of Chernozhukov et al. (2007), which they impose directly: $\|\Phi_v(\theta, P)\| \geq C \cdot [d(\theta, \Theta_I) \wedge \delta]$ on Θ for some $C, \delta > 0$. Together with Assumption 5.1, this gives their Condition C.2 with $\gamma = 1$; see the proof of Proposition 5.3.

Under these assumptions, the set estimator $\widehat{\Theta}_I(\tau)$ converges to Θ_I in the Hausdorff metric; the results are stated formally as follows.

Proposition 5.3 (Consistency and convergence rate of $\widehat{\Theta}_I$)**.** *Suppose that Assumptions 4.9 and 5.1 hold with $\Theta_I = \Theta_I(\Phi, P)$ defined in (4). Let $\tau \in \mathbb{R}_+$ be such that $\tau \geq \sup_{\theta \in \Theta_I} \sqrt{n} \cdot Q_n(\theta)$ with probability approaching 1 and $\tau = o_P(\sqrt{n})$. Then $d_H(\widehat{\Theta}_I, \Theta_I) = o_P(1)$; more precisely,*

$$d_H(\widehat{\Theta}_I, \Theta_I) = O_P\left((1 \vee \tau)/\sqrt{n} \right).$$

If, in addition, Assumption 5.2 holds, then one can set $\tau = \sqrt{n} \cdot c_n$ and obtain

$$d_H(\widehat{\Theta}_I, \Theta_I) = O_P\left(1/\sqrt{n} \right).$$

Proof. See Appendix B.5. □

[20] Yildiz (2012) instead put conditions on moment functions; see Kaido et al. (2022) for their relationships in terms of constraint qualifications.

5.2. **Limiting distribution of $\widehat{\Theta}_I$.** Now I focus on the estimator $\widehat{\Theta}_I = \widehat{\Theta}_I(\tau)$ with $\tau = \sqrt{n} \cdot c_n$ and investigate its limiting distribution measured in the Hausdorff metric. In this section, I focus on moment inequality restrictions, i.e. $|\mathcal{J}_2| = d_q$, as this is the prototypical case in the literature. The key device underlying my analysis is the stochastic field whose supremum represents the Hausdorff metric between the estimator $\widehat{\Theta}_I$ and Θ_I, i.e.,

$$Z_n(x) := \sqrt{n} \cdot \left(d(x, \widehat{\Theta}_I) - d(x, \Theta_I) \right). \tag{25}$$

By a property of d_H (e.g. Rockafellar and Wets (1998, Example 4.13)), $\sup_{x \in \mathbb{R}^{d_\theta}} |Z_n(x)| = \sqrt{n} \cdot d_H(\widehat{\Theta}_I, \Theta_I)$. This stochastic field is not equicontinuous; a similar issue arises in Chernozhukov et al. (2007) for the QLR statistic $\sup_{\theta' \in \Theta_I} a_n \cdot Q_n(\theta')$, which they use to construct asymptotically valid confidence sets. Hence care is needed in handling Z_n. As in Chernozhukov et al. (2007), I decompose $\partial\Theta_I$ by the set of binding restrictions, which makes the limit a Lipschitz functional of the Gaussian process D, so that the continuous mapping theorem applies.

In deriving the limiting distribution of $\widehat{\Theta}_I$, I add the following assumptions.

Assumption 5.4. (i). $\Phi(\cdot, P) : \Theta \mapsto \mathbb{R}^{d_q}$ is C^2 on $\Theta \subset \mathbb{R}^{d_\theta}$. (ii). There exists a constant $c > 0$ such that, for any $\theta \in \partial\Theta_I$,

$$d\left(0,\ \text{conv}\left\{\phi_j(\theta) : j \in \mathcal{A}(\theta)\right\}\right) \geq c > 0,$$

where $\phi_j(\theta) := \nabla_\theta \Phi_j(\theta, P)$ for $j \in \mathcal{J}_2$ and $\mathcal{A}(\theta) := \mathcal{A}(\theta, P)$ is the set of binding restrictions in Assumption 4.9. (iii). $\{\phi_j(\theta)\}_{j \in \mathcal{A}(\theta)}$ are linearly independent at any $\theta \in \partial\Theta_I$. (iv). $\Theta_I \subset \Theta^o$.

Assumption 5.4 (i) is a smoothness condition on $\Phi(\cdot, P)$, which makes my argument simple in this section. Note that this is a condition on $\Phi(\cdot, P)$ and it does not exclude sets like polyhedrons. The set $\mathcal{A}(\theta)$ captures all binding moment restrictions. Assumption 5.4 (ii) and (iii) are related to the Mangasarian-Fromovitz Constraint Qualification (MFCQ) and the LICQ, respectively, which are often used in the literature in partially identified models; see Kaido et al. (2022) for the detailed discussion.

Remark 5.5. It is possible to consider the other cases such as $|\mathcal{J}_1| = d_q$ or $|\mathcal{J}_1| + |\mathcal{J}_2| = d_q$. The first case nests the point identified situation, while the second case corresponds to the one, where for example a parameter is point identified via moment equalities and the moment inequalities impose additional restrictions. Each has its own difficulty. A difficulty in the first case, for example, arises when the parameter is under identified; under identified non-linear IV case corresponds to this situation as mentioned in Chernozhukov et al. (2007). In this case, even $\sqrt{n}$ consistency is not established in the criterion function based approach. A difficulty in the second case occurs when a parameter is uniquely identified via moment equalities and moment inequalities bind at the same time; this situation requires specific boundary-corrected asymptotic analysis as in Andrews (1999). A detailed investigation of these cases is left for future research.

My main result here shows that the (scaled) Hausdorff-based distribution of $\widehat{\Theta}_I$, which is centered at the target set Θ_I, weakly converges to the supremum of a Gaussian field on $\partial\Theta_I$. The formal statement is as follows.

Theorem 5.6 (Limiting distribution of $\widehat{\Theta}_I$ in the Hausdorff metric.)**.** *Suppose that Assumptions 4.9, 5.1, 5.2, and 5.4 hold. Then for the set estimator $\widehat{\Theta}_I(\tau)$ with $\tau = \sqrt{n} \cdot c_n$ with $c_n = \inf_{\theta' \in \Theta} Q_n(\theta')$ it holds that*

$$\begin{aligned} &\sqrt{n} \cdot d_H(\widehat{\Theta}_I, \Theta_I) \\ &\quad \rightsquigarrow \sup_{\theta \in \partial\Theta_I} \sup \left\{ |\alpha^t D_{\mathcal{A}(\theta)}(\theta)| : \alpha \geq 0, \left\| \alpha^t G_{\mathcal{A}(\theta)}(\theta) \right\| = 1 \right\}, \quad \text{as } n \to \infty, \end{aligned} \tag{26}$$

where $D_{\mathcal{A}(\theta)}(\theta)$ and $G_{\mathcal{A}(\theta)}(\theta)$ consist of the rows $j \in \mathcal{A}(\theta)$ of

$$D(\theta) := \left(\nabla q_j(Pm(\theta, \cdot))\, \xi(\theta) \right)_{j \in \mathcal{J}_2} \in \mathbb{R}^{|\mathcal{J}_2|} \quad \text{and} \quad G(\theta) := \nabla_\theta \Phi(\theta, P) \in \mathbb{R}^{|\mathcal{J}_2| \times d_\theta},$$

respectively, with $\xi(\theta)$ in (24), i.e.

$$\begin{aligned} D_{\mathcal{A}(\theta)}(\theta) &= (D_j(\theta))_{j \in \mathcal{A}(\theta)} = \left(\nabla q_j(Pm(\theta, \cdot))\, \xi(\theta) \right)_{j \in \mathcal{A}(\theta)}, \\ G_{\mathcal{A}(\theta)}(\theta) &= (G_j(\theta))_{j \in \mathcal{A}(\theta)} = (\phi_j(\theta)^t)_{j \in \mathcal{A}(\theta)}. \end{aligned} \tag{27}$$

Proof. See Appendix B.6. □

Theorem 5.6 has a simple local reading. Near $\theta \in \partial\Theta_I$, the identified set looks like θ plus the cone $T(\theta) = \{\psi : G_{\mathcal{A}(\theta)}(\theta)\psi \leq 0\}$, and the estimator $\widehat{\Theta}_I$ looks like the same cone shifted by $v/\sqrt{n}$, where $G_{\mathcal{A}(\theta)}(\theta)v = -D_{\mathcal{A}(\theta)}(\theta)$: replacing each binding restriction by its sample analogue moves the corresponding face of Θ_I along its normal. The Hausdorff distance between a cone and its translate is the largest component of the shift along a unit normal $u = G_{\mathcal{A}(\theta)}(\theta)^t\alpha$ with $\alpha \geq 0$, which is the inner supremum in (26). At a smooth boundary point ($|\mathcal{A}(\theta)| = 1$), it equals $|D_j(\theta)|/\|\phi_j(\theta)\|$, the noise in the binding restriction divided by its slope. At an edge or a vertex ($|\mathcal{A}(\theta)| > 1$), the shifts of the faces combine, and the vertex can move further than any single face. The outer supremum then picks the boundary point where this normal displacement is largest; interior points of Θ_I play no role.

5.3. **Hausdorff-ball confidence set.** The result in Section 5.2 suggests a bootstrap-based confidence set, which is similar to the one for the convex identified set (Beresteanu and Molinari, 2008; Kaido, 2016) and differs from the approach taken in much of the literature, where confidence sets are constructed as level sets of the criterion function (e.g., Chernozhukov et al., 2007; Bugni, 2010; Romano and Shaikh, 2010; Chen et al., 2018). Motivated by the result in Section 5.2, I here briefly illustrate how to compute the bootstrap-based confidence set. As in Section 5.2, I focus on the case of moment inequalities. The validity of this bootstrap procedure is left for future research.

The key observation is that the set estimator $\widehat{\Theta}_I$ has the following relation

$$\Theta_I \subset (\widehat{\Theta}_I)^r \quad \Leftrightarrow \quad \overrightarrow{d_H}(\Theta_I, \widehat{\Theta}_I) \leq r,$$

for $r > 0$, where $d_H^{\rightarrow}(\Theta_I, \widehat{\Theta}_I) = \sup_{\theta \in \Theta_I} d(\theta, \widehat{\Theta}_I)$ is the directed Hausdorff distance. Note that $(\widehat{\Theta}_I)^r$ is the closed r-envelope of $\widehat{\Theta}_I$; see Appendix A.2. To construct a valid confidence set, the problem is to find an $r > 0$ with suitable coverage property. This quantity can be derived from

$$C_n(\alpha) := (\widehat{\Theta}_I)^{r_n^{\rightarrow}(\alpha)}, \quad r_n^{\rightarrow}(\alpha) = n^{-1/2} \cdot \zeta_n^{\rightarrow}(1-\alpha),$$

where $\alpha \in (0,1)$ and $\zeta_n^{\rightarrow}(1-\alpha)$ is the $(1-\alpha)$ quantile of the conditional distribution, given the data, of

$$D_n^{\rightarrow *} := \sqrt{n} \cdot d_H^{\rightarrow}(\widehat{\Theta}_I, \widehat{\Theta}_I^*),$$

with $\widehat{\Theta}_I^*$ denoting the bootstrap set estimator specified below. By the definition of the envelope, $\Theta_I \subset C_n(\alpha)$ if and only if $\sqrt{n} \cdot d_H^{\rightarrow}(\Theta_I, \widehat{\Theta}_I) \leq \zeta_n^{\rightarrow}(1-\alpha)$. Hence, if the conditional distribution of $D_n^{\rightarrow *}$ consistently estimates the limiting distribution of $\sqrt{n} \cdot d_H^{\rightarrow}(\Theta_I, \widehat{\Theta}_I)$, and the latter is continuous at its $(1-\alpha)$ quantile, then

$$\lim_{n \to \infty} P\{\Theta_I \subset C_n(\alpha)\} = 1 - \alpha,$$

so that $C_n(\alpha)$ has asymptotic coverage $1-\alpha$. A formal proof of the validity of the bootstrap is beyond the scope of this paper.

For the construction of $\widehat{\Theta}_I^*$, I consider the standard procedure. Let $e_1, \cdots, e_n$ be i.i.d. with mean zero and unit variance and independent of the data (e.g., $e_i \sim N(0,1)$), and let

$$\mathbb{P}_n^* m(\theta) := \mathbb{P}_n m(\theta) + \frac{1}{n} \sum_{i=1}^n e_i \cdot \{m(\theta, X_i) - \mathbb{P}_n m(\theta)\},$$
$$Q_n^*(\theta) := \left\| \Phi_v^t(\theta, \mathbb{P}_n^*) \cdot W_n^{1/2} \right\|, \quad c_n^* := \inf_{\theta' \in \Theta} Q_n^*(\theta').$$

The bootstrap set estimator $\widehat{\Theta}_I^*$ is defined as

$$\widehat{\Theta}_I^* := \{\theta \in \Theta : Q_n^*(\theta) \leq c_n^*\}.$$

The algorithm for computing the confidence set $C_n(\alpha) = (\widehat{\Theta}_I)^{r_n^{\rightarrow}(\alpha)}$ is illustrated in Algorithm 1.

Remark 5.7 (Two-sided version). The confidence set $C_n(\alpha)$ is one-sided: it controls only how far Θ_I can extend beyond $\widehat{\Theta}_I$. A two-sided version is obtained by replacing the directed distance $d_H^{\rightarrow}$ with the Hausdorff distance d_H throughout. Let $\zeta_n(1-\alpha)$ be the $(1-\alpha)$ quantile of the conditional distribution, given the data, of

$$D_n^* := \sqrt{n} \cdot d_H(\widehat{\Theta}_I, \widehat{\Theta}_I^*),$$

and let $r_n(\alpha) := n^{-1/2} \cdot \zeta_n(1-\alpha)$. Since $d_H(\Theta_I, \widehat{\Theta}_I) \leq r$ if and only if both $\Theta_I \subset (\widehat{\Theta}_I)^r$ and $\widehat{\Theta}_I \subset (\Theta_I)^r$, the same argument as above gives, under the analogous conditions,

$$\lim_{n \to \infty} P\{d_H(\Theta_I, \widehat{\Theta}_I) \leq r_n(\alpha)\} = 1 - \alpha.$$

That is, with asymptotic probability $1-\alpha$, the identified set lies in the Hausdorff ball of radius $r_n(\alpha)$ centred at $\widehat{\Theta}_I$: Θ_I is contained in $(\widehat{\Theta}_I)^{r_n(\alpha)}$, and in addition no point of

$\widehat{\Theta}_I$ is farther than $r_n(\alpha)$ from Θ_I. Because $d_H \geq d_H^{\rightarrow}$, the radius satisfies $r_n(\alpha) \geq r_n^{\rightarrow}(\alpha)$; this is the price of the additional statement. In Algorithm 1, only the bootstrap statistic changes, to

$$D_n^{(b)} := \sqrt{n} \cdot \max\Big\{ \max_{\theta_g \in \widehat{\Theta}_{I,s}} \delta_g^{(b)},\ \max_{\theta_g \in \widehat{\Theta}_{I,s}^{(b)}} \hat{\delta}_g \Big\},$$

with ζ_n and r_n in place of $\zeta_n^{\rightarrow}$ and $r_n^{\rightarrow}$.

Algorithm 1 One-sided Hausdorff-ball confidence set (grid implementation)

Grid and precomputation

Form the grid $\Theta_s := \Theta \cap \big(\theta_s + s\,\mathbb{Z}^{d_\theta}\big) = \{\theta_1, \cdots, \theta_G\}$.

For $g = 1, \cdots, G$ and $i = 1, \cdots, n$: $\bar{m}_g := \mathbb{P}_n m(\theta_g)$ and $a_{gi} := m(\theta_g, X_i) - \bar{m}_g$.

Set estimate

$Q_g := \big\|\Phi_v^t(\theta_g, \mathbb{P}_n) \cdot W_n^{1/2}\big\|$ for all g, where $\Phi(\theta_g, \mathbb{P}_n) = q(\bar{m}_g)$, and $\hat{c}_n := \min_g Q_g$.

$\widehat{\Theta}_{I,s} := \{\theta_g :\ Q_g \leq \hat{c}_n\}$, and $\hat{\delta}_g := d(\theta_g, \widehat{\Theta}_{I,s})$ for all g.

Bootstrap

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

  draw $e_1^{(b)}, \cdots, e_n^{(b)}$ i.i.d. from the multiplier law (e.g., $N(0,1)$)

  $\bar{m}_g^{(b)} := \bar{m}_g + n^{-1} \sum_{i=1}^n e_i^{(b)} a_{gi}$ and $Q_g^{(b)} := \big\|\Phi_v^t(\theta_g, \mathbb{P}_n^{(b)}) \cdot W_n^{1/2}\big\|$ for all g, where $\Phi(\theta_g, \mathbb{P}_n^{(b)}) = q(\bar{m}_g^{(b)})$

  $\hat{c}_n^{(b)} := \min_g Q_g^{(b)}$ and $\widehat{\Theta}_{I,s}^{(b)} := \{\theta_g :\ Q_g^{(b)} \leq \hat{c}_n^{(b)}\}$

  $\delta_g^{(b)} := d\big(\theta_g, \widehat{\Theta}_{I,s}^{(b)}\big)$ for all g

  $D_n^{\rightarrow(b)} := \sqrt{n}\, \max_{\theta_g \in \widehat{\Theta}_{I,s}} \delta_g^{(b)}$

Critical value and output

$\zeta_n^{\rightarrow} := D_n^{\rightarrow(k)}$, the k-th smallest of $D_n^{\rightarrow(1)}, \cdots, D_n^{\rightarrow(B)}$, with $k := \lceil (B+1)(1-\alpha) \rceil$; set $r_n^{\rightarrow} := \zeta_n^{\rightarrow}/\sqrt{n}$.

$C_{n,s}(\alpha) := \{\theta_g \in \Theta_s :\ \hat{\delta}_g \leq r_n^{\rightarrow}\}$.

6. Finite sample properties

6.1. **Data-generating process.** I here evaluate the finite-sample performance of the set estimator $\widehat{\Theta}_I$ and the Hausdorff-ball confidence set $C_n(\alpha)$ in Section 5.3, using the optimal top tax rate example in Section 2.

Model. For convenience, I briefly recap the model in Section 2. Taxpayers with income above a threshold z^* face a constant marginal rate on income above z^* and misperceive the tax schedule as in Farhi and Gabaix (2020): a taxpayer with income z responds to a weighted average of the true marginal rate, with attention weight $\omega \in [0,1]$, and

of a weighted average of the marginal rates at other income levels. In the revenue-maximizing case without income effects, the optimal top rate is

$$\bar{\tau} = \frac{1}{1 + \bar{\zeta}_{c,r} \cdot \pi \cdot \kappa(A)}, \quad \kappa(A) = \omega + (1-\omega)A, \tag{28}$$

where $\bar{\zeta}_{c,r}$ is the elasticity of a taxpayer who perceives the schedule correctly, $\pi = z_m/(z_m - z^*)$ is the Pareto parameter of the upper tail with $z_m = \mathbb{E}_P[Z \mid Z \geq z^*]$ and Z reported income, and A represents the shape of misperception. At $\kappa = 1$, i.e., $\omega = 1$ or $A = 1$, (28) is the formula of Saez (2001). The basic conditions maintained through this exercise are as follows. The attention weight $\omega = \omega_0$ and the elasticity $\bar{\zeta}_{c,r}$ are assumed to be known, the misperception shape lies in a maintained range $A \in [A_L, A_U]$, and the tail parameter is partially identified because incomes are topcoded at a known cap $z_c > z^*$. Under the setting, the identified set of our interest for the parameter $\theta = (\bar{\tau}, A)^t$ is expressed as (see Proposition 3.6 for detail)

$$\Theta_I = \left\{ (\bar{\tau}, A)^t \in \Theta : \; A_L \leq A \leq A_U, \; \frac{1}{1 + \bar{\zeta}_{c,r}\pi_U \kappa(A)} \leq \bar{\tau} \leq \frac{1}{1 + \bar{\zeta}_{c,r}\pi_L \kappa(A)} \right\}, \tag{29}$$

where, with $S = P(Z \geq z^*)$ and the censoring moments

$$\begin{aligned} M_L &= \mathbb{E}_P\left[(Z \wedge z_c) \cdot \mathbb{1}\{Z \geq z^*\}\right], \\ M_U &= \mathbb{E}_P\left[(Z \cdot \mathbb{1}\{Z < z_c\} + \bar{z} \cdot \mathbb{1}\{Z \geq z_c\}) \cdot \mathbb{1}\{Z \geq z^*\}\right], \end{aligned}$$

the bounds on the tail parameter are expressed

$$\pi_L = \frac{M_U/S}{M_U/S - z^*}, \quad \pi_U = \frac{M_L/S}{M_L/S - z^*}.$$

Note that $\bar{z}$ is the maintained upper bound on the mean income above the cap in Assumption 3.5(ii).

Below I explain the detail of the data generating process for ingredients of the identified set in (29).

Population. I normalize the bracket threshold to $z^* = 1$ and let reported income in the top bracket follow an exact Pareto law with tail index π_0, i.e.,

$$P(Z > t) = t^{-\pi_0}, \quad t \geq 1, \quad \pi_0 = 1.5, \tag{30}$$

where the value $\pi_0 = 1.5$ is consistent with the stability of the Pareto parameter above roughly \$300,000 documented in Diamond and Saez (2011). Since every object in Θ_I depends on the income distribution only through moments conditional on $Z \geq z^*$, I simulate the top-bracket sample directly; the sampling design is therefore $Z_1, \cdots, Z_n$ i.i.d. from (30), so that $S = P(Z \geq z^*) = 1$ in this sample[21].

Censoring. Recall that reported income is topcoded at a known cap z_c. Accordingly, the econometrician observes

$$(Z_i \wedge z_c, \; D_i), \quad D_i = \mathbb{1}\{Z_i \geq z_c\}, \quad i \in [n].$$

[21] With a full-population sample, S is estimated as well and the only change is an additional moment, which is strongly correlated with the two moments M_L and M_U.

The cap z_c is set to the $(1-p_c)$ quantile of (30), i.e., $z_c = p_c^{-1/\pi_0}$, where p_c is the share of the bracket that is censored. I use $p_c = 0.05$, which gives $z_c = 7.37$, and set $\bar{z} = 50$. Assumption 3.5(ii) holds, since under (30) the mean income above the cap is $\pi_0 z_c/(\pi_0 - 1) = 22.1$; the interval for π below contains π_0 precisely because of this.

Implied bounds on the tail parameter. Under (30) the censoring moments are available in closed form. With the density $f(z) = \pi_0 z^{-\pi_0 - 1}$, it follows that

$$M_L = \mathbb{E}_P[Z \wedge z_c] = \frac{\pi_0}{\pi_0 - 1}\left(1 - z_c^{-(\pi_0 - 1)}\right) + z_c \cdot p_c,$$
$$M_U = \mathbb{E}_P\left[Z \cdot \mathbb{1}\{Z < z_c\} + \bar{z} \cdot \mathbb{1}\{Z \geq z_c\}\right] = M_L + (\bar{z} - z_c) \cdot p_c.$$

At $p_c = 0.05$ these equal $M_L = 2.2632$ and $M_U = 4.3948$. Hence the expressions for π_L and π_U imply

$$\pi \in [\pi_L, \pi_U] = [1.2946,\ 1.7916], \tag{31}$$

which contains the true value $\pi_0 = 1.5$.

Behavioral parameters. I consider two values of the attention weight, $\omega_0 \in \{0.2,\ 0.45\}$, and set the range of the misperception shape to $[A_L, A_U] = [0.55, 1.3]$. The two values are close to the range of estimates of attention to the French income tax in Lardeux (2023), who finds that tax filers assign a weight of 0.22 to 0.44 to the effective tax schedule, depending on the estimation method. The range also contains the estimate of attention to sales taxes in Taubinsky and Rees-Jones (2018), 0.25. The lower endpoint A_L is the value implied by ironing, $A = 1/\pi$, at the largest admissible tail parameter $\pi_U = 1.79$; Rees-Jones and Taubinsky (2020) find that taxpayers commonly iron. The upper endpoint allows for moderate over-perception of the top rate, for which I am not aware of a direct estimate. The true misperception shape need not be specified, since Θ_I depends only on the range.

Sources of the calibration. The tail index $\pi_0 = 1.5$ and the elasticity $\bar{\zeta}_{c,r} = 0.25$ are the values used in Diamond and Saez (2011); the elasticity $\bar{\zeta}_{c,r} = 0.57$ is the estimate for taxable income at the top of the distribution in Gruber and Saez (2002), which Diamond and Saez (2011) regard as a conservative upper bound.

Designs. I consider four designs, depending on the values of $\omega_0 \in \{0.2, 0.45\}$ and $\bar{\zeta}_{c,r} = \{0.25, 0.57\}$; for other parameters, all of the designs share $\pi_0 = 1.5$, $p_c = 0.05$, $\bar{z} = 50$ and $[A_L, A_U] = [0.55, 1.3]$. Figure 2 displays Θ_I in the four designs. The section at $A = 1$ is the interval implied by Saez (2001) at $e = \bar{\zeta}_{c,r}$ and does not depend on ω_0; points with $A < 1$ lie above it and points with $A > 1$ below it. A lower attention weight makes the optimal rate more sensitive to A and the upper boundary more curved.

For each design, I consider the sample sizes $n \in \{500,\ 1{,}000,\ 5{,}000\}$ with 1,000 Monte Carlo replications and 499 bootstrap or subsampling draws per replication.

6.2. **Results.** As set estimators and confidence sets, here I compare different methods described below.

(i) *Hausdorff ball.* The confidence set $C_n(\alpha)$ of Section 5.3: the set estimator $\widehat{\Theta}_I$ enlarged by the radius $r_n^{\rightarrow}(\alpha) = n^{-1/2}\zeta_n^{\rightarrow}(1-\alpha)$, where $\zeta_n^{\rightarrow}(1-\alpha)$ is the

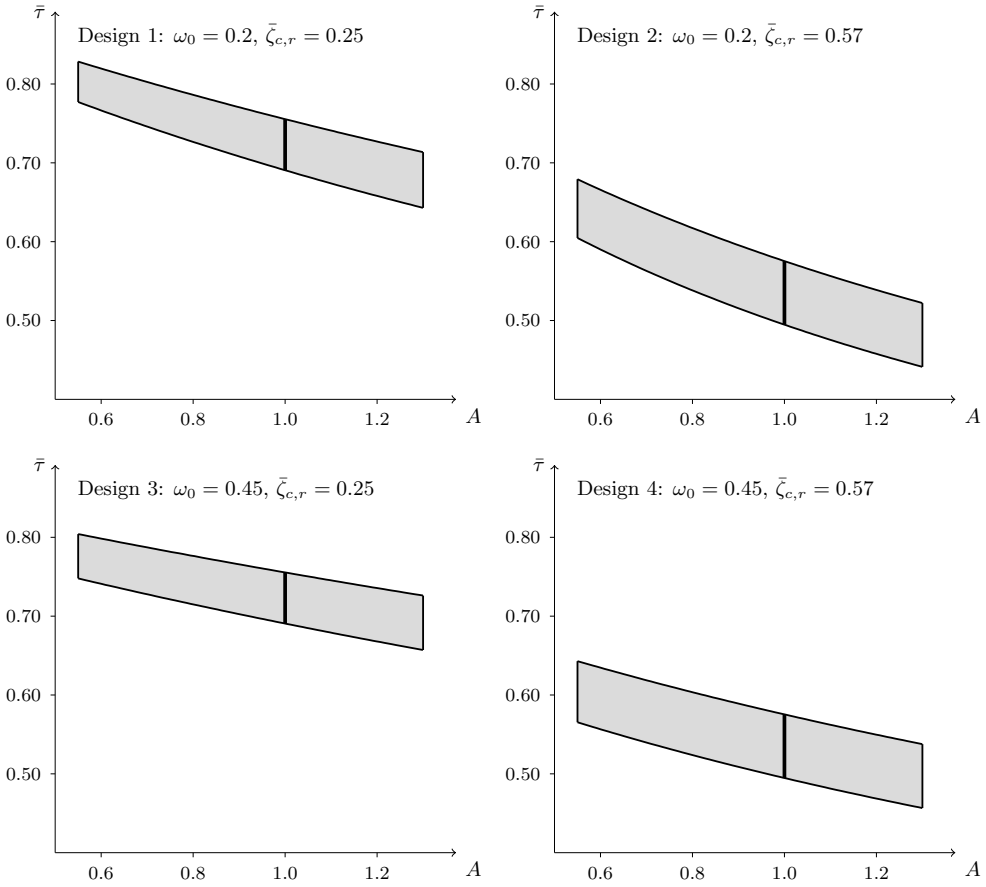


FIGURE 2. The population identified set (29) in the four designs. The thick vertical segment is the section at $A = 1$.

$(1-\alpha)$ quantile of the multiplier bootstrap distribution of the one-sided distance $\sqrt{n} \cdot d_H^{\rightarrow}$, as in Algorithm 1. In this section $C_n(\alpha)$ is intersected with the known range $A_L \leq A \leq A_U$ (The other four methods impose this range as the non-stochastic constraints $\Phi_3, \Phi_4 \leq 0$).

(ii) *Support function.* The Wald-type confidence set of Kaido (2016), which builds on the support-function approach of Beresteanu and Molinari (2008) and Kaido and Santos (2014). I use the subsampling version as implemented in Kaido (2016).[22]

(iii) *Subsampling.* The confidence set of Chernozhukov et al. (2007): a level set of the criterion function, with the critical value obtained by subsampling the supremum of the criterion over the estimated set. I use a single iteration started from $\widehat{\Theta}_I$, as their Condition C.3 permits.

[22]The upper face of $\text{conv}(\Theta_I)$ is a straight segment whose slope depends on P, which is excluded in Kaido and Santos (2014) (Remark 3.2 and Assumption 4.1) and makes a bootstrap of the support-function statistic doubtful; subsampling requires only a continuous limit distribution.

(iv) *Bootstrap.* The bootstrap of Bugni (2010), which approximates the distribution of the same supremum by retaining only the moment inequalities that are close to binding, with the threshold $\tau_n = \ln \ln n$.

(v) *Monte Carlo confidence set.* Procedure 2 of Chen et al. (2018). The moment inequalities are written as moment equalities with slackness parameters, quasi-posterior draws are generated by their sequential Monte Carlo sampler with 10,000 draws, and the confidence set consists of the values of θ whose profiled quasi-likelihood ratio lies below the $(1-\alpha)$ quantile of its draws.

For the two subsampling methods, (ii) and (iii), the subsample size is $b = \lfloor 0.2n \rfloor$, which follows the rule $b_n \propto n$ of Bugni (2010). All resampling methods use $B = 499$ draws, and their critical values are the $\lceil (B+1)(1-\alpha) \rceil$-th order statistic of these draws.

Set estimator. Table 1 reports the median Hausdorff distance of $\widehat{\Theta}_I$ and of the support-function estimator $\mathrm{conv}(\widehat{\Theta}_I)$ to Θ_I. The distances of $\widehat{\Theta}_I$ are smaller than thoes of $\mathrm{conv}(\widehat{\Theta}_I)$, which furthermore falls much more slowly as the sample size increases. Its ratio to the distance of $\widehat{\Theta}_I$ therefore grows with n, up to 5.00 in design 2 at $n = 5{,}000$. Figure 3 compares the distribution of $\sqrt{n} \cdot d_H(\widehat{\Theta}_I, \Theta_I)$ with the limiting distribution of Theorem 5.6. At $n = 1{,}000$ the two are almost indistinguishable in all four designs.

				$\widehat{\Theta}_I$			Support function			Ratio		
Design	ω_0	$\bar{\zeta}_{c,r}$	$n =$	500	1,000	5,000	500	1,000	5,000	500	1,000	5,000
1	0.2	0.25		0.55	0.39	0.18	0.70	0.58	0.43	1.27	1.47	2.43
2	0.2	0.57		0.66	0.46	0.20	1.16	1.05	1.00	1.75	2.27	5.00
3	0.45	0.25		0.55	0.40	0.18	0.61	0.46	0.26	1.11	1.14	1.47
4	0.45	0.57		0.67	0.46	0.20	0.83	0.65	0.48	1.25	1.43	2.42

TABLE 1. Median Hausdorff distance to Θ_I of the set estimators, times 100.

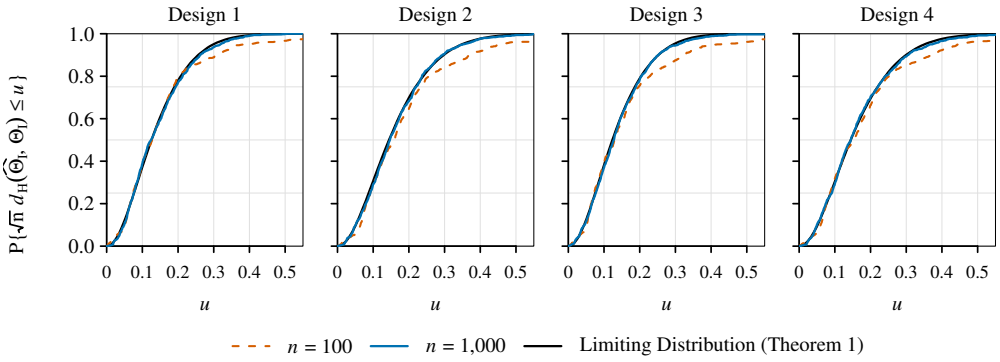


FIGURE 3. Distribution function of $\sqrt{n}\, d_H(\widehat{\Theta}_I, \Theta_I)$ (at $n = 100$ and $n = 1{,}000$) and the limiting distribution of Theorem 5.6.

Coverage and Hausdorff distance. Panel A and B of Table 2 report coverage and median Hausdorff distance to Θ_I at the nominal level 0.90. The Hausdorff ball covers Θ_I with

frequency 0.908–0.932 at $n = 500$ and 0.893–0.906 at $n = 5{,}000$. The bootstrap of Bugni (2010) and the Monte Carlo confidence set of Chen et al. (2018) under-cover slightly at $n = 500$, with 0.871–0.907, and are close to the nominal level at $n = 5{,}000$, with 0.891–0.900. The subsampling confidence set of Chernozhukov et al. (2007) has coverage that falls with n, to 0.863–0.871 at $n = 5{,}000$. With b/n held fixed, this is in line with the finding of Bugni (2010). The support-function confidence set is conservative, with coverage 0.96–0.99, as expected for a set that targets $\text{conv}(\Theta_I)$. For Hausdorff distance, methods except the support function have similar distance at $n = 5,000$. The similar results can be observed for the case of the nominal level 0.95 in Table 3.

				Panel A. Coverage			**Panel B.** $100\times$ **median** d_H		
Design	ω_0	$\bar{\zeta}_{c,r}$	Method $n =$	500	1,000	5,000	500	1,000	5,000
1	0.2	0.25	Hausdorff ball $C_n(\alpha)$	0.932	0.909	0.906	1.49	1.04	0.45
			Support function	0.985	0.969	0.963	2.78	2.38	1.92
			Chernozhukov et al. (2007), subsampling	0.927	0.876	0.863	1.35	0.94	0.41
			Bugni (2010), bootstrap	0.905	0.888	0.897	1.30	0.95	0.44
			Chen et al. (2018)	0.907	0.889	0.898	1.32	0.96	0.44
2	0.2	0.57	Hausdorff ball $C_n(\alpha)$	0.908	0.924	0.902	1.73	1.24	0.52
			Support function	0.968	0.978	0.964	3.22	2.77	2.22
			Chernozhukov et al. (2007), subsampling	0.907	0.898	0.871	1.57	1.09	0.47
			Bugni (2010), bootstrap	0.878	0.903	0.898	1.51	1.12	0.50
			Chen et al. (2018)	0.882	0.906	0.897	1.52	1.12	0.50
3	0.45	0.25	Hausdorff ball $C_n(\alpha)$	0.915	0.911	0.902	1.50	1.03	0.45
			Support function	0.977	0.968	0.963	2.57	2.15	1.65
			Chernozhukov et al. (2007), subsampling	0.913	0.881	0.863	1.38	0.92	0.41
			Bugni (2010), bootstrap	0.871	0.890	0.900	1.31	0.93	0.43
			Chen et al. (2018)	0.880	0.896	0.900	1.33	0.94	0.43
4	0.45	0.57	Hausdorff ball $C_n(\alpha)$	0.915	0.905	0.893	1.78	1.21	0.52
			Support function	0.972	0.962	0.970	2.80	2.23	1.60
			Chernozhukov et al. (2007), subsampling	0.907	0.880	0.870	1.61	1.08	0.48
			Bugni (2010), bootstrap	0.886	0.881	0.891	1.55	1.10	0.50
			Chen et al. (2018)	0.891	0.887	0.897	1.57	1.11	0.51

TABLE 2. Coverage of Θ_I by the 90% confidence sets (Panel A) and their median Hausdorff distance to Θ_I, times 100 (Panel B).

				Panel A. Coverage			**Panel B.** $100\times$ **median** d_H		
Design	ω_0	$\bar{\zeta}_{c,r}$	Method $n=$	500	1,000	5,000	500	1,000	5,000
1	0.2	0.25	Hausdorff ball $C_n(\alpha)$	0.981	0.963	0.953	1.76	1.22	0.52
			Support function	0.994	0.989	0.987	3.06	2.53	1.97
			Chernozhukov et al. (2007), subsampling	0.969	0.949	0.933	1.60	1.08	0.47
			Bugni (2010), bootstrap	0.950	0.936	0.946	1.47	1.08	0.49
			Chen et al. (2018)	0.951	0.938	0.947	1.48	1.09	0.50
2	0.2	0.57	Hausdorff ball $C_n(\alpha)$	0.965	0.978	0.956	2.10	1.45	0.61
			Support function	0.987	0.991	0.985	3.52	2.94	2.28
			Chernozhukov et al. (2007), subsampling	0.958	0.961	0.929	1.89	1.28	0.54
			Bugni (2010), bootstrap	0.932	0.955	0.948	1.70	1.26	0.57
			Chen et al. (2018)	0.939	0.954	0.950	1.72	1.26	0.57
3	0.45	0.25	Hausdorff ball $C_n(\alpha)$	0.974	0.962	0.952	1.77	1.20	0.52
			Support function	0.993	0.986	0.980	2.83	2.29	1.71
			Chernozhukov et al. (2007), subsampling	0.971	0.941	0.921	1.64	1.07	0.46
			Bugni (2010), bootstrap	0.924	0.939	0.947	1.48	1.05	0.49
			Chen et al. (2018)	0.938	0.944	0.945	1.51	1.06	0.49
4	0.45	0.57	Hausdorff ball $C_n(\alpha)$	0.966	0.958	0.963	2.13	1.43	0.61
			Support function	0.992	0.984	0.990	3.13	2.41	1.66
			Chernozhukov et al. (2007), subsampling	0.964	0.934	0.936	1.92	1.26	0.55
			Bugni (2010), bootstrap	0.930	0.929	0.952	1.74	1.23	0.57
			Chen et al. (2018)	0.935	0.933	0.956	1.75	1.24	0.57

TABLE 3. Coverage of Θ_I by the 95% confidence sets (Panel A) and their median Hausdorff distance to Θ_I, times 100 (Panel B).

7. DISCUSSION

This paper studies the asymptotic efficiency of set estimation for a general class of identified sets. I frist derive the asymptotic minimax efficiency bound for set estimators under risk functions based on the Hausdorff metric. The analysis relies on a limit experiment framework tailored to the local structure of the identified set, and the bound recovers the classical efficiency bound as a special case when the parameter is point identified. I then show that a set estimator based on a criterion function is efficient under mild conditions. To establish this, I derive the limiting distribution of the set estimator in the Hausdorff metric without globally characterizing the identified set through its support functions, which allows the identified set to be non-convex. This limiting distribution motivates a bootstrap-based confidence set for the identified set and I examine its finite sample performance. I illustrate the framework with an application to optimal taxation with behavioral agents.

Appendices

A. Some primers

A.1. **Notation.** For any $r \in [1, \infty]$, $\|.\|_{L^r}$ denotes the L^r-norm. For any vector $a \in \mathbb{R}^d$, $\|.\|$ denotes the Euclidean norm. For positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, let $a_n \lesssim b_n$ if $\limsup_n a_n/b_n < \infty$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For any set A, $|A|$ represents the cardinality of A. For a set A in a topological space, $\overline{A}$, A^o, and ∂A represent the closure, the interior, and the boundary of A respectively. For a metric space (V, d_V), $B_v(\epsilon)$ is an open ball centered at $v \in V$ with radius $\epsilon > 0$. For a and b in $\mathbb{R}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a vector $u \in \mathbb{R}^d$, let $u^{\otimes 2} = uu^T$ denote the outer product. Given a square matrix A, $A > 0$ and $A < 0$ indicate that A is positive definite (p.d.) and negative definite (n.d.) respectively; positive semi-definite (p.s.d.) and negative semi-definite (n.s.d.) correspond to weak inequalities. For any random variable X, its law is denoted by $\mathcal{L}(X)$. Given a probability measure P and a measurable function f, let $\mathbb{E}_P f = \int f dP$. For any function f, $\mathrm{Im}(f)$ represents the image of f. For sets A and B, $A \oplus B$ represents their Minkowski sum; see Appendix A.2 for the definition.

A.2. **Random set theory.** Throughout this section, $\mathcal{F}, \mathcal{G}$ and $\mathcal{K}$ represent the family of non-empty closed, open and compact sets in $\mathbb{R}^{d_\theta}$ respectively.

Definition A.1 (Basic set operations)**.** For any $A, B \subset \mathbb{R}^{d_\theta}$ and a scalar α,

$$A \oplus B := \{a + b : \ a \in A, b \in B\}, \quad \alpha A := \{\alpha a : \ a \in A\}.$$

The operation $A \oplus B$ is called the *Minkowski sum* of sets A and B.

Definition A.2 (Hausdorff metric d_H)**.** For any $A, B \in \mathcal{K}$,

$$d_H(A, B) := \max\left\{\sup_{a\in A}\inf_{b\in B}\|a - b\|, \ \sup_{b\in B}\inf_{a\in A}\|a - b\|\right\}. \tag{A.1}$$

The Hausdorff metric d_H admits alternative representations (Molchanov (2017) Proposition D.8). The first one connects d_H to ϵ-envelop of sets; for a set A, its ϵ-envelop denoted by A^ϵ is the set

$$A^\epsilon := \{x \in \mathbb{R}^{d_\theta} : \ \|x - A\| \le \epsilon\}. \tag{A.2}$$

Then d_H is equivalent to

$$d_H(A, B) = \max\left\{\rho_H(A, B), \ \rho_H(B, A)\right\}, \quad A, B \in \mathcal{K}, \tag{A.3}$$

where $\rho_H(A, B) := \inf\{\epsilon > 0 : \ A \subset B^\epsilon\}$. The second one takes the form

$$d_H(A, B) = \sup_{x\in\mathbb{R}^{d_\theta}} \left|\|x - A\| - \|x - B\|\right|. \tag{A.4}$$

Note that this formulation works in general for $A, B \in \mathcal{F}$ as well with the possibility that $d_H(A, B) = \infty$ (Rockafellar and Wets (1998) Chapter 4).

Definition A.3 (Myopic topology on $\mathcal{K}$)**.** The *myopic* topology on $\mathcal{K}$ has a sub-base consisting of

$$\begin{aligned} \mathcal{K}^F &= \{K \in \mathcal{K} : K \cap F = \emptyset\}, \quad F \in \mathcal{F}, \\ \mathcal{K}_G &= \{K \in \mathcal{K} : K \cap G \neq \emptyset\}, \quad G \in \mathcal{G}. \end{aligned} \tag{A.5}$$

Under the condition that the underlying metric space is separable, which corresponds to the present case $\mathbb{R}^{d_\theta}$, the following two types of sets are open in the topology induced by the Hausdorff metric: the sets $\{K \in \mathcal{K} : K \subset G\}$ and $\{K \in \mathcal{K} : K \cap G \neq \emptyset\}$ for $G \in \mathcal{G}$ (e.g. Theorem D.6. in Molchanov (2017)). In particular, this implies that the Borel σ-algebra $\mathcal{B}_\mathcal{K}$ generated by the Hausdorff metric is equivalent to the one generated by the Myopic topology.

Definition A.4 (Random compact set)**.** A map $T : \mathcal{X} \mapsto \mathcal{K}$ defined on a probability space $(\mathcal{X}, \mathcal{B}_\mathcal{X}, P)$ is called a *random compact set* if

$$\{x \in \mathcal{X} : \ T(x) \cap K \neq \emptyset\} \in \mathcal{B}_\mathcal{X} \quad \text{for any } K \in \mathcal{K}.$$

More generally, a map $T : \mathcal{X} \mapsto \mathcal{F}(\mathcal{G})$ is called a *random closed (open) set*. For the present purpose, it suffices to focus on random compact sets. Note that the above definition is equivalent to defining a random compact set T as a measurable map from $(\mathcal{X}, \mathcal{B}_\mathcal{X}, P)$ to $(\mathcal{K}, \mathcal{B}_\mathcal{K})$, where $\mathcal{B}_\mathcal{K}$ is the Borel σ-field induced by the Hausdorff metric (see Molchanov (2017) section 1.3.2 and Appendix D for the detail).

Definition A.5 (Containment functional)**.** Given a random compact set T, the functional $C_T : \mathcal{K} \mapsto [0, 1]$ such that

$$C_T(K) = \mathbb{P}(T \subset K), \quad K \in \mathcal{K},$$

is called the *containment functional*.

B. Theoretical background

B.1. **Proofs for Section 3.3.**

Proof. Proposition 3.6. I start from (a). First note that since $z_c > z^*$, every $Q \in \mathcal{Q}(P)$ satisfies $Q(Z \geq z^*) = S$. On the event $\{D = 0\}$ we have $Z = \tilde{Z}$ while on the event $\{D = 1\}$ we have $Z \geq z_c$ and $\tilde{Z} = z_c \geq z^*$. This implies $\mathbb{E}_Q[Z \cdot \mathbb{1}\{Z \geq z_c\}] = P(D = 1) \cdot \mathbb{E}_Q[Z \,|\, Z \geq z_c] \in [z_c \cdot P(D = 1), \bar{z} \cdot P(D = 1)]$. Hence

$$\mathbb{E}_Q\left[Z \cdot \mathbb{1}\{Z \geq z^*\}\right] = \mathbb{E}_P\left[\tilde{Z}(1 - D) \cdot \mathbb{1}\{\tilde{Z} \geq z^*\}\right] + \mathbb{E}_Q\left[Z \cdot \mathbb{1}\{Z \geq z_c\}\right] \in [M_L, M_U],$$

so that $z_{m,Q} \in [M_L/S, M_U/S]$. Conversely, for $\lambda \in [0, 1]$, let Q_λ be such that it agrees with P on $\{Z < z_c\}$ and that, conditional on $\{Z \geq z_c\}$, it puts mass $1 - \lambda$ at z_c and λ at $\bar{z}$. Then $Q_\lambda \in \mathcal{Q}(P)$, since $\mathbb{E}_{Q_\lambda}[Z \,|\, Z \geq z_c] = z_c + \lambda(\bar{z} - z_c) \leq \bar{z}$, and $z_{m,Q_\lambda} = (M_L + \lambda(M_U - M_L))/S$, which attains every point of $[M_L/S, M_U/S]$. By Assumption 3.5(iv), $M_L/S > z^*$, and by Assumption 3.5(ii) and (iv), $M_U > M_L$ with $M_U/S < \infty$. The map $z_m \mapsto z_m/(z_m - z^*)$ is continuous and strictly decreasing on (z^*, ∞) with values in $(1, \infty)$, so it maps $[M_L/S, M_U/S]$ onto $[\pi_L, \pi_U]$ with $1 < \pi_L < \pi_U < \infty$. This shows the result in (a).

For (b), note that the result in (a) implies that $\theta \in \Theta_I$ if and only if $A_L \leq A \leq A_U$ and $\pi_L \leq \pi(\theta) \leq \pi_U$. Since $M_U - z^*S > 0$ and $M_L - z^*S > 0$, multiplying through by these quantities gives

$$\begin{aligned}
\pi(\theta) \geq \pi_L &\iff \pi(\theta)(M_U - z^*S) \geq M_U \iff \Phi_1(\theta, P) \leq 0, \\
\pi(\theta) \leq \pi_U &\iff \pi(\theta)(M_L - z^*S) \leq M_L \iff \Phi_2(\theta, P) \leq 0,
\end{aligned}$$

and $A_L \leq A \leq A_U$ if and only if $\Phi_3, \Phi_4 \leq 0$. Finally, $\pi(\theta)$ is strictly decreasing in $\bar{\tau}$ for each A, and $\pi(\theta) = \pi$ if and only if $\bar{\tau} = 1/(1 + \bar{\zeta}_{c,r}\pi\kappa(A))$, so $\pi_L \leq \pi(\theta) \leq \pi_U$ is the pair of inequalities on $\bar{\tau}$ in (3). □

B.2. **Proofs for Section 4.1.**

B.2.1. *Proofs for the main text.* I first state and prove the expansion of the log-likelihood ratios used in Section 4.1.

Lemma B.1 (Local asymptotic normality along the paths)**.** *Suppose that Assumption 4.2 (ii) holds. Then for any $\theta \in \Theta_I$ and $h \in \mathbb{R}^{d_\theta}$, it holds that $P_0 \dot{\ell}_{\theta,h} h = 0$ and $\sigma_\theta(h) = P_0(\dot{\ell}_{\theta,h} h)^2 < \infty$. Moreover, for any finite collection of pairs $(\theta_j, h_j) \in \Theta_I \times \mathbb{R}^{d_\theta}$, $j = 1, \cdots, J$, it holds under P_0^n that*

$$\log \frac{dP^n_{\theta_j + h_j/\sqrt{n}}}{dP^n_0} = \mathbb{G}_n \left(\dot{\ell}_{\theta_j, h_j} h_j \right) - \frac{1}{2} \sigma_{\theta_j}(h_j) + o_{P_0}(1), \quad j = 1, \cdots, J,$$

where $\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P_0)$ is the empirical process with $\mathbb{P}_n$ the empirical measure, that

$$\left(\log \frac{dP^n_{\theta_j + h_j/\sqrt{n}}}{dP^n_0} \right)_{j=1}^J \rightsquigarrow N_J \left(-\frac{1}{2} \left(\sigma_{\theta_j}(h_j) \right)_{j=1}^J , \ \left(P_0 \, \dot{\ell}_{\theta_j, h_j} h_j \, \dot{\ell}_{\theta_k, h_k} h_k \right)_{j,k=1}^J \right) \quad \text{as } n \to \infty,$$

and that $P^n_{\theta_j + h_j/\sqrt{n}}$ and P^n_0 are mutually contiguous for each $j = 1, \cdots, J$.

Proof. Lemma B.1. Fix $\theta \in \Theta_I$ and $h \in \mathbb{R}^{d_\theta}$, and write $p_n := p_{\theta + h/\sqrt{n}}$ and $s := \dot{\ell}_{\theta,h} h$. Since $p_\theta = p_0$ μ-a.e., Assumption 4.2 (ii) states that $\sqrt{n}(\sqrt{p_n} - \sqrt{p_0}) \to \frac{1}{2} s \sqrt{p_0}$ in $L^2(\mu)$, which also implies $\sqrt{p_n} \to \sqrt{p_0}$ in $L^2(\mu)$. Since both p_n and p_0 are probability densities, it holds that

$$\begin{aligned} 0 = \sqrt{n} \int (p_n - p_0) \, d\mu &= \int \sqrt{n} \left(\sqrt{p_n} - \sqrt{p_0} \right) \left(\sqrt{p_n} + \sqrt{p_0} \right) d\mu \\ &\overset{(a)}{\to} \int \frac{1}{2} s \sqrt{p_0} \cdot 2 \sqrt{p_0} \, d\mu = P_0 s, \end{aligned}$$

where (a) follows from the continuity of the inner product in $L^2(\mu)$. Hence $P_0 s = 0$, and $\sigma_\theta(h) = P_0 s^2 < \infty$ as $s \in L^2(P_0)$. The proof of Theorem 7.2 in van der Vaart (1998) uses the differentiability in quadratic mean only along the sequence $\theta + h/\sqrt{n}$, together with $P_0 s = 0$; the same steps under Assumption 4.2 (ii) provide the expansion for each pair (θ_j, h_j). By the multivariate central limit theorem, $(\mathbb{G}_n(\dot{\ell}_{\theta_j, h_j} h_j))_{j=1}^J$ converges in law under P_0^n to the centered normal distribution with the covariance matrix $(P_0 \, \dot{\ell}_{\theta_j, h_j} h_j \, \dot{\ell}_{\theta_k, h_k} h_k)_{j,k=1}^J$; combining this with the expansion and Slutsky's lemma provides the joint convergence. Finally, each marginal limit is $N(-\frac{1}{2}\sigma_{\theta_j}(h_j), \sigma_{\theta_j}(h_j))$, whose mean equals minus half the variance; Example 6.5 in van der Vaart (1998) then shows that $P^n_{\theta_j + h_j/\sqrt{n}}$ and P^n_0 are mutually contiguous. □

Proof. Lemma 4.3. Let $r > 0$ be as in the condition. Since $T(\vartheta)$ and $N(\vartheta)$ in (11) are the tangent and normal cones of Federer (1959) (Definitions 4.3 and 4.4 there), Theorem 4.8 (12) there shows that every $u \in N(\vartheta)$ with $\|u\| = 1$ is of the form $v/\|v\|$

with ϑ the unique nearest point in Θ_I of $\vartheta + v$. Theorem 4.8 (7) there, applied to $\vartheta + v$, then gives

$$u^t(z - \vartheta) \le \frac{\|z - \vartheta\|^2}{2r} \quad \text{for all } \vartheta, z \in \Theta_I \text{ and } u \in N(\vartheta) \text{ with } \|u\| \le 1. \tag{B.1}$$

For $\vartheta \in \Theta_I$ and $w \in \mathbb{R}^{d_\theta}$, let

$$\sigma_\vartheta(w) := \sup\left\{u^t w : u \in N(\vartheta),\ \|u\| \le 1\right\}.$$

It is non-negative and 1-Lipschitz in w. By (17), for $\Delta \in \mathcal{C}(\Theta_I)$,

$$\|\Delta\|_\perp = \sup_{\vartheta \in \Theta_I} \max\left\{\sigma_\vartheta(\Delta(\vartheta)),\ \sigma_\vartheta(-\Delta(\vartheta))\right\}. \tag{B.2}$$

The idea of the proof is as follows. Each of the two directed Hausdorff distances between the moved sets is a supremum, over $\vartheta \in \Theta_I$, of the distance from a moved point to the other moved set. I show that, to the first order, this distance depends only on the relative displacement $g - g'$: it equals the distance from $\vartheta + (g - g')(\vartheta)/\sqrt{n}$ to Θ_I itself. I then evaluate the latter distance by positive reach. It is $1/\sqrt{n}$ times the normal component of $(g - g')(\vartheta)$, up to a term of order $1/n$.

Fix a compact set $\mathcal{C}_0 \subset \mathcal{C}(\Theta_I)$. By the Arzelà–Ascoli theorem, the maps in $\mathcal{C}_0$ are uniformly bounded and uniformly equicontinuous on Θ_I. Let $c := 2\sup_{g \in \mathcal{C}_0} \|g\|_\infty < \infty$, and let ω be a common modulus of uniform continuity of the maps in $\mathcal{C}_0$. Fix $g, g' \in \mathcal{C}_0$ and write $t := 1/\sqrt{n}$. Let $K_n := \Theta_I \oplus_t g$ and $K_n' := \Theta_I \oplus_t g'$, and let $\Delta := g - g'$. Then $\|\Delta\|_\infty \le c$, and Δ has the modulus 2ω. For $y \in \mathbb{R}^{d_\theta}$ and a compact set K, write $d(y, K) := \min_{z \in K} \|y - z\|$.

The directed distance from K_n to K_n' is $\sup_{\vartheta \in \Theta_I} d(\vartheta + tg(\vartheta), K_n')$. I show that

$$\lim_{n \to \infty} \sup_{g, g' \in \mathcal{C}_0} \left| \sqrt{n} \sup_{\vartheta \in \Theta_I} d\left(\vartheta + tg(\vartheta), K_n'\right) - \sup_{\vartheta \in \Theta_I} \sigma_\vartheta(\Delta(\vartheta)) \right| = 0. \tag{B.3}$$

This provides the result. Indeed, exchanging g and g' replaces Δ by $-\Delta$. Hence $\sqrt{n} \cdot d_H(K_n, K_n')$, the maximum of the two directed distances, converges uniformly over $g, g' \in \mathcal{C}_0$ to the right-hand side of (B.2), i.e., to $\|\Delta\|_\perp$.

The first step is to show that only the relative displacement Δ matters. Fix $\vartheta \in \Theta_I$. Since $K_n' = \{z + tg'(z) : z \in \Theta_I\}$,

$$d\left(\vartheta + tg(\vartheta), K_n'\right) = \min_{z \in \Theta_I} \left\|\vartheta + t\Delta(\vartheta) - z + t\left(g'(\vartheta) - g'(z)\right)\right\|.$$

Taking $z = \vartheta$ shows that this minimum is at most $t\|\Delta(\vartheta)\| \le ct$. On the other hand, the expression in the minimum is at least $\|\vartheta - z\| - ct$. Hence every minimizer z satisfies $\|z - \vartheta\| \le 2ct$. In the same way, every nearest point z of Θ_I to $\vartheta + t\Delta(\vartheta)$ satisfies $\|z - \vartheta\| \le 2ct$. For such z, $t\|g'(\vartheta) - g'(z)\| \le t\,\omega(2ct)$. Restricting both minimizations to these z therefore gives

$$\left|d\left(\vartheta + tg(\vartheta), K_n'\right) - d\left(\vartheta + t\Delta(\vartheta), \Theta_I\right)\right| \le t\,\omega(2ct), \quad \vartheta \in \Theta_I,\ g, g' \in \mathcal{C}_0. \tag{B.4}$$

It remains to evaluate $d(\vartheta + t\Delta(\vartheta), \Theta_I)$, starting with a lower bound. Let z be a nearest point of Θ_I to $\vartheta + t\Delta(\vartheta)$, and let $u \in N(\vartheta)$ with $\|u\| \le 1$. By (B.1) and

$\|z - \vartheta\| \le 2ct$,

$$d\left(\vartheta + t\Delta(\vartheta), \Theta_I\right) \ge u^t\left(\vartheta + t\Delta(\vartheta) - z\right)$$
$$= t\, u^t \Delta(\vartheta) - u^t(z - \vartheta) \ge t\, u^t \Delta(\vartheta) - \frac{2c^2t^2}{r}.$$

Taking the supremum over u gives

$$d\left(\vartheta + t\Delta(\vartheta), \Theta_I\right) \ge t\, \sigma_\vartheta(\Delta(\vartheta)) - \frac{2c^2t^2}{r}. \tag{B.5}$$

For the upper bound, let $y := \vartheta + t\Delta(\vartheta)$, and let p be a nearest point of Θ_I to y, so that $\|p - \vartheta\| \le 2ct$. Suppose first that $y \ne p$. Then $u := (y - p)/\|y - p\|$ lies in $N(p)$. To see this, let $h \in T(p)$, with $p + t_k h_k \in \Theta_I$, $t_k \searrow 0$ and $h_k \to h$. Since p is a nearest point, $\|y - p - t_k h_k\|^2 \ge \|y - p\|^2$, i.e., $-2(y - p)^t h_k + t_k \|h_k\|^2 \ge 0$. Letting $k \to \infty$ gives $(y - p)^t h \le 0$. Hence, by (B.1) with the roles of ϑ and p exchanged,

$$d(y, \Theta_I) = u^t(y - p) = t\, u^t \Delta(\vartheta) + u^t(\vartheta - p) \le t\, \sigma_p(\Delta(\vartheta)) + \frac{\|\vartheta - p\|^2}{2r}$$
$$\le t\, \sigma_p(\Delta(p)) + 2t\, \omega(2ct) + \frac{2c^2t^2}{r}.$$

The last inequality uses that σ_p is 1-Lipschitz and $\|\Delta(\vartheta) - \Delta(p)\| \le 2\omega(2ct)$. If $y = p$, the same bound holds trivially, as $\sigma_p \ge 0$. Hence

$$d\left(\vartheta + t\Delta(\vartheta), \Theta_I\right) \le t \sup_{\vartheta' \in \Theta_I} \sigma_{\vartheta'}(\Delta(\vartheta')) + 2t\, \omega(2ct) + \frac{2c^2t^2}{r}. \tag{B.6}$$

Finally, I combine the bounds. Take the supremum over $\vartheta \in \Theta_I$ in (B.5) and (B.6), and multiply by $\sqrt{n} = 1/t$. For bounded functions a and b on Θ_I,

$$\left| \sup_{\vartheta \in \Theta_I} a(\vartheta) - \sup_{\vartheta \in \Theta_I} b(\vartheta) \right| \le \sup_{\vartheta \in \Theta_I} |a(\vartheta) - b(\vartheta)|.$$

Together with (B.4), this gives

$$\left| \sqrt{n} \sup_{\vartheta \in \Theta_I} d\left(\vartheta + tg(\vartheta), K_n'\right) - \sup_{\vartheta \in \Theta_I} \sigma_\vartheta(\Delta(\vartheta)) \right| \le 3\, \omega\left(\frac{2c}{\sqrt{n}}\right) + \frac{2c^2}{r\sqrt{n}}$$

for all $g, g' \in \mathcal{C}_0$, which shows (B.3). □

Proof. Theorem 4.4. The proof extends that of Theorem 7.10 in van der Vaart (1998). Since $\hat{g}_n$ takes values in $\mathcal{C}(\Theta_I)$ rather than in a Euclidean space, the representation step uses Theorem 3.1 in van der Vaart (1991). A last step passes from $\hat{g}_n$ to the Hausdorff distance by Lemma 4.3.

Fix a finite set $D \subset \Theta_I \times \mathbb{R}^{d_\theta}$. For $(\theta, h) \in D$, write $P_{n,\theta,h} := P^n_{\theta + h/\sqrt{n}}$, $L_{n,\theta,h} := \log(dP_{n,\theta,h}/dP_0^n)$ and $\Theta^{\theta,h}_{I,n} := \Theta_I(P_{\theta+h/\sqrt{n}})$. Recall that $\mathcal{C}(\Theta_I)$ equipped with $\|\cdot\|_\infty$ is a separable Banach space, hence Polish.

I first show the joint convergence under P_0. By the condition (ii) and Lemma B.1, the random element $(\hat{g}_n, (L_{n,\theta,h})_{(\theta,h)\in D})$ is uniformly tight in $\mathcal{C}(\Theta_I) \times \mathbb{R}^{|D|}$ under P_0^n. By Prokhorov's theorem, there exists a subsequence $\{n_k\}_{k\in\mathbb{N}}$ along which this random

element converges in law under $P_0^{n_k}$. By Lemma B.1, the limit of the second component is the normal distribution given there.

These limits are next represented in the limit experiment. By Le Cam's third lemma (Theorem 3.11.7 in van der Vaart and Wellner (2023)), together with this convergence and the contiguity in Lemma B.1, $\hat{g}_{n_k}$ converges in law under $P_{n_k,\theta,h}$ to some probability measure $\mathcal{L}_{\theta,h}$ on $\mathcal{C}(\Theta_I)$, for each $(\theta,h) \in D$. Moreover, by Lemma B.1, the experiments $\{P_{n,\theta,h} : (\theta,h) \in D\}$ converge to the experiments $\{G_s : s = \dot{\ell}_{\theta,h}h,\ (\theta,h) \in D\}$ in (16), which are dominated by G_0. Lemma B.3 and Theorem 3.1 in van der Vaart (1991) then provide a randomized statistic $T = T(\mathcal{G}, U)$ with values in $\mathcal{C}(\Theta_I)$, where U is uniformly distributed on $[0,1]$ and independent of $\mathcal{G}$. For each $(\theta,h) \in D$, $\mathcal{L}_{\theta,h}$ is the law of T under $\mathcal{G} \sim G_s$ with $s = \dot{\ell}_{\theta,h}h$.

It remains to pass from $\hat{g}_n$ to the Hausdorff distance. Fix $(\theta,h) \in D$ and write $g := g_{\theta,h}$. For $z \in \mathcal{C}(\Theta_I)$, let

$$f_n(z) := \sqrt{n} \cdot d_H\left(\Theta_I \oplus_{1/\sqrt{n}} z,\ \Theta_I \oplus_{1/\sqrt{n}} g\right), \quad f(z) := \|z - g\|_\perp.$$

The map f is continuous, as $\|\cdot\|_\perp \leq \|\cdot\|_\infty$. By the triangle inequality for d_H, it holds that

$$\begin{aligned} &\left|\sqrt{n} \cdot d_H\left(T_n, \Theta_{I,n}^{\theta,h}\right) - f_n(\hat{g}_n)\right| \\ &\quad \leq \sqrt{n} \cdot d_H\left(T_n, \Theta_I \oplus_{1/\sqrt{n}} \hat{g}_n\right) + \sqrt{n} \cdot d_H\left(\Theta_{I,n}^{\theta,h}, \Theta_I \oplus_{1/\sqrt{n}} g\right). \end{aligned} \tag{B.7}$$

The first term on the right-hand side converges to 0 in P_0^n-probability by the condition (ii). By the contiguity in Lemma B.1, it then also converges to 0 in $P_{n,\theta,h}$-probability. The second term converges to 0 by Assumption 4.2 (iii). If $z_n \to z$ in $\mathcal{C}(\Theta_I)$, then

$$|f_n(z_n) - f(z)| \leq |f_n(z_n) - f(z_n)| + |f(z_n) - f(z)| \overset{(a)}{\to} 0,$$

where (a) follows from Lemma 4.3 applied to the compact set $\{z_n\}_{n\in\mathbb{N}} \cup \{z, g\}$ and the continuity of f. The extended continuous mapping theorem (Theorem 18.11 in van der Vaart (1998)) and the representation above then give $f_{n_k}(\hat{g}_{n_k}) \rightsquigarrow f(T) = \|T - g\|_\perp$ under $P_{n_k,\theta,h}$, where $\mathcal{G} \sim G_s$ with $s = \dot{\ell}_{\theta,h}h$. Combining this with (B.7) and Slutsky's lemma provides the result. □

B.2.2. *Auxiliary lemmas.* I here state specific notions and results used for the main results in Section 4.1. Let $\mathcal{B}_\mathcal{C}$ be the Borel σ-field of $\mathcal{C}(\Theta_I)$ equipped with $\|\cdot\|_\infty$.

Definition B.2 (Markov kernel)**.** A *Markov kernel* $\delta : \mathcal{C}(\Theta_I) \times \mathcal{X} \mapsto \mathbb{R}$ defined in $(\mathcal{C}(\Theta_I) \times \mathcal{X}, \mathcal{B}_\mathcal{C} \otimes \mathcal{B}_\mathcal{X})$ is such that

(i). for any $x \in \mathcal{X}$, $\delta(.,x)$ is a probability measure on $(\mathcal{C}(\Theta_I), \mathcal{B}_\mathcal{C})$,
(ii). for any $A \in \mathcal{B}_\mathcal{C}$, the map $\delta(A,.) : \mathcal{X} \mapsto \mathbb{R}$ is measurable.

Lemma B.3. *The following relations hold for a random statistic T and a Markov kernel δ:*

(i). *For any Markov kernel* $\delta : \mathcal{C}(\Theta_I) \times \mathcal{X} \mapsto \mathbb{R}$, *there exists a measurable map* $T : (\mathcal{X} \times [0,1], \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{[0,1]}) \mapsto (\mathcal{C}(\Theta_I), \mathcal{B}_{\mathcal{C}})$ *such that*

$$\delta(A, x) = \mathbb{P}(T(x, U) \in A), \quad \text{for any } A \in \mathcal{B}_{\mathcal{C}}, \tag{B.8}$$

where U *is uniformly distributed on* $[0,1]$.

(ii). *Given a map* T *above, the equation (B.8) defines a Markov kernel* $\delta : \mathcal{C}(\Theta_I) \times \mathcal{X} \mapsto \mathbb{R}$.

Proof. This is the result in Lemma 9.1 in van der Vaart (1991) with $(\mathcal{C}(\Theta_I), \mathcal{B}_{\mathcal{C}})$ the target space, noting that $\mathcal{C}(\Theta_I)$ equipped with $\|\cdot\|_\infty$ is a separable Banach space, hence Polish, as Θ_I is compact. □

The next lemma is used in the proof of Lemma B.5 below: it shows that a set in $\mathcal{K}^*_\Theta$ with non-empty interior is locally defined by finitely many inequalities with linearly independent gradients. This is a standard consequence of the inverse function theorem; for a set with smooth boundary, i.e., without corners, see Theorem 5.48 in Lee (2012), where such sets are called regular domains.

Lemma B.4. *Let* $K \in \mathcal{K}_2$, *or* $K \in \mathcal{K}_3$ *with* $k = d_\theta$. *Then for every* $x \in \partial K$ *there exist an open neighborhood* $U_x \subset \mathbb{R}^{d_\theta}$ *of* x, *an integer* $k_x \geq 1$ *and infinitely differentiable functions* $\psi_{x,1}, \cdots, \psi_{x,k_x} : U_x \mapsto \mathbb{R}$ *such that* $\psi_{x,i}(x) = 0$ *for* $i = 1, \cdots, k_x$, *the gradients* $\nabla\psi_{x,1}(x), \cdots, \nabla\psi_{x,k_x}(x)$ *are linearly independent, and* $K \cap U_x = \{y \in U_x : \psi_{x,i}(y) \leq 0,\ i = 1, \cdots, k_x\}$.

Proof.

Case 1: $K \in \mathcal{K}_2$. Here $d_\theta = 1$ and K is a finite union of disjoint compact intervals with non-empty interior. A point $x \in \partial K$ is an endpoint of exactly one of these intervals, say $[a, b]$ with $a < b$. Take $k_x = 1$, an open interval U_x around x that does not meet the other intervals and does not contain the other endpoint, and $\psi_{x,1}(y) := a - y$ if $x = a$ and $\psi_{x,1}(y) := y - b$ if $x = b$.

Case 2: $K \in \mathcal{K}_3$ *with* $k = d_\theta$. Fix $x \in \partial K$ and let (V, φ) be a chart for K with $x \in V$, either an interior chart or a chart with corners, and let $z := \varphi(x)$. Write $E := \mathbb{R}^{d_\theta}$ for an interior chart and $E := \overline{\mathbb{R}}^{d_\theta}_+$ for a chart with corners, so that $\varphi(V) = O \cap E$ for some open set $O \subset \mathbb{R}^{d_\theta}$. Since the inclusion map ι_K is smooth, $\iota_K \circ \varphi^{-1}$ admits a C^∞ extension F to an open neighborhood $W \subset O$ of z, and since ι_K is an immersion, the differential $DF(z)$ is injective, hence invertible as $k = d_\theta$. By the inverse function theorem, I may shrink W so that F is a C^∞ diffeomorphism from W onto the open set $F(W)$. Since V is open in K, $V = K \cap O'$ for some open set $O' \subset \mathbb{R}^{d_\theta}$, and by the continuity of φ there exists an open neighborhood $Q \subset O'$ of x with $\varphi(K \cap Q) \subset W$. Let $U := Q \cap F(W)$, an open neighborhood of $x = F(z)$. If $y \in K \cap U$, then $y \in V$ and $w := \varphi(y) \in W \cap E$ satisfies $F(w) = \varphi^{-1}(w) = y$, so that $F^{-1}(y) \in E$. Conversely, if $y \in U$ and $w := F^{-1}(y) \in E$, then $w \in W \cap E \subset \varphi(V)$, so that $y = F(w) = \varphi^{-1}(w) \in K$. Hence

$$K \cap U = \left\{y \in U : F^{-1}(y) \in E\right\}. \tag{B.9}$$

If $E = \mathbb{R}^{d_\theta}$, (B.9) gives $U \subset K$, contradicting $x \in \partial K$; hence (V, φ) is a chart with corners. Let $I := \{i : z^i = 0\}$. If $I = \emptyset$, then z is an interior point of $\overline{\mathbb{R}}^{d_\theta}_+$, and shrinking

U so that $F^{-1}(U)$ lies in the interior of $\overline{\mathbb{R}}_+^{d_\theta}$ gives $U \subset K$, again a contradiction; hence $k_x := |I| \geq 1$. By continuity, I may shrink U so that $(F^{-1}(y))^i > 0$ for every $y \in U$ and $i \notin I$. Then (B.9) reads $K \cap U = \{y \in U : -(F^{-1}(y))^i \leq 0,\ i \in I\}$. Let $U_x := U$ and let $\psi_{x,1}, \cdots, \psi_{x,k_x}$ be the functions $-(F^{-1})^i|_U$, $i \in I$. These are infinitely differentiable, vanish at x as $z^i = 0$ for $i \in I$, and their gradients at x are, up to sign, rows of $D(F^{-1})(x) = DF(z)^{-1}$, an invertible matrix, hence linearly independent. □

The next lemma collects the sets with positive reach used in the paper; see the paragraph after Lemma 4.3.

Lemma B.5. *Let $K \in \mathcal{K}_\Theta$. Then K has positive reach in each of the following cases:*

(a). *K is finite;*
(b). *K is a finite union of disjoint compact convex sets;*
(c). *for every $x \in K$ there exist an open neighborhood V of x, disjoint finite index sets $\mathcal{J}_1$ and $\mathcal{J}_2$ and functions $f_j : V \mapsto \mathbb{R}$ of class C^2, $j \in \mathcal{J}_1 \cup \mathcal{J}_2$, such that $K \cap V = \{y \in V : f_j(y) = 0,\ j \in \mathcal{J}_1,\ f_j(y) \leq 0,\ j \in \mathcal{J}_2\}$ and the gradients $\nabla f_j(x)$, $j \in \mathcal{J}_1 \cup \{j \in \mathcal{J}_2 : f_j(x) = 0\}$, are linearly independent.*

In particular, every $K \in \mathcal{K}_1 \cup \mathcal{K}_2$, and every $K \in \mathcal{K}_3$ that is either full-dimensional or has an empty manifold boundary, has positive reach.

Proof.

(a) If $K = \{\theta_1, \cdots, \theta_J\}$ and $\delta := \min_{i \neq j} \|\theta_i - \theta_j\|$, every point within distance $\delta/2$ of K has a unique nearest point in K.

(b) Let $K = \bigcup_i K_i$ with disjoint compact convex sets K_i, and let $\delta := \min_{i \neq j} \min\{\|y - z\| : y \in K_i,\ z \in K_j\} > 0$. If y is within distance $\delta/2$ of K, all its nearest points in K lie in a single K_i, since nearest points in K_i and K_j with $i \neq j$ would be within distance δ of each other; and a point has a unique nearest point in a closed convex set.

(c) Fix $x \in K$. With $F := (f_j)_{j \in \mathcal{J}_1 \cup \mathcal{J}_2}$ and $D := \{0\}^{|\mathcal{J}_1|} \times (-\infty, 0]^{|\mathcal{J}_2|}$, it holds that $K \cap V = \{y \in V : F(y) \in D\}$, and the linear independence of the binding gradients implies the constraint qualification in Definition 10.23 of Rockafellar and Wets (1998); hence K is strongly amenable at x (Definition 10.23 (b) there), and so is its indicator function (the remark after that definition). By Proposition 13.32 and Exercise 13.31 there, K is prox-regular at x, and by Theorem 1.3, (a) and (k), in Poliquin et al. (2000), the projection onto K is then single-valued on a neighborhood of x, which I may take to contain the open ball of radius $2\varepsilon_x$ around x for some $\varepsilon_x > 0$. Since K is compact, it is covered by finitely many open balls of radius ε_{x_l} around points $x_l \in K$, $l = 1, \cdots, L$; let $\varepsilon := \min_l \varepsilon_{x_l} > 0$. If y is within distance ε of K and p is a nearest point of y in K, then $\|p - x_l\| < \varepsilon_{x_l}$ for some l, so that $\|y - x_l\| < 2\varepsilon_{x_l}$ and y has a unique nearest point in K.

Finally, the sets in $\mathcal{K}_1$ and $\mathcal{K}_2$ are covered by (a) and (b), the full-dimensional sets in $\mathcal{K}_3$ satisfy (c) by Lemma B.4, and a compact embedded smooth submanifold of dimension $k < d_\theta$ without boundary is locally the zero set of $d_\theta - k$ smooth functions with linearly independent gradients (e.g. Lee (2012, Chapter 5)), so that it satisfies (c). □

B.3. **Proofs for Section 4.2.**

B.3.1. *Proofs for the main text.*

Proof. Theorem 4.6. The proof extends that of Theorem 8.11 in van der Vaart (1998) for the point identified case, in three parts. First, Theorem 4.4 reduces the asymptotic risk to the limit experiment. Second, the minimax risk there is bounded below by Bayes risks with normal priors, as in Proposition 8.6 in van der Vaart (1998); Anderson's lemma used there is replaced by Lemma B.6, since T need not lie in the range of Υ. Third, the bound is extended from bounded continuous loss functions to general ones.

Denote the left hand side of the inequality in Theorem 4.6 by $\mathcal{R}(l)$. Let $s(a)$ and $R(T,a)$ be as in Lemma B.7 below. By Assumption 4.5, for each $a \in \mathbb{R}^{d_a}$ there exists $(\theta_a, h_a) \in \Theta_I \times \mathbb{R}^{d_\theta}$ such that $\dot{\ell}_{\theta_a,h_a} h_a = s(a)$. Then $g_{\theta_a,h_a} = \Upsilon a$.

I first reduce the asymptotic risk to the limit experiment when l is bounded, uniformly continuous and non-decreasing. Fix a finite set $A \subset \mathbb{R}^{d_a}$ and let $D_A := \{(\theta_a, h_a) : a \in A\} \in \mathcal{D}$. There exists a subsequence $\{n_m\}_{m\in\mathbb{N}}$ along which

$$\max_{a\in A} \mathbb{E}_{P^n_{\theta_a+h_a/\sqrt{n}}} \left[l \circ \left(\sqrt{n} \cdot d_H \left(T_n, \Theta_I(P_{\theta_a+h_a/\sqrt{n}}) \right) \right) \right]$$

converges to its limit inferior. The proof of Theorem 4.4 applies without change along $\{n_m\}_{m\in\mathbb{N}}$. It provides a further subsequence and a randomized statistic $T_A = T_A(\mathcal{G}, U)$ with values in $\mathcal{C}(\Theta_I)$ such that, for each $a \in A$, along this subsequence,

$$\sqrt{n} \cdot d_H \left(T_n, \Theta_I(P_{\theta_a+h_a/\sqrt{n}}) \right) \rightsquigarrow \|T_A - \Upsilon a\|_\perp,$$

under $P^n_{\theta_a+h_a/\sqrt{n}}$ on the left-hand side and $\mathcal{G} \sim G_{s(a)}$ on the right-hand side. Since l is bounded and continuous, it follows that

$$\mathcal{R}(l) \geq \max_{a\in A} R(T_A, a) \geq \inf_T \max_{a\in A} R(T, a), \tag{B.10}$$

where the infimum is taken over all randomized statistics $T = T(\mathcal{G}, U)$ with values in $\mathcal{C}(\Theta_I)$.

The next step passes from finite sets to a Bayes risk, for l as above with modulus of continuity ω_l. For any T and $a, b \in \mathbb{R}^{d_a}$, it holds that

$$|R(T,a) - R(T,b)| \overset{(a)}{\leq} \sup |l| \cdot \mathbb{E}_{G_0} \left| \frac{dG_{s(a)}}{dG_0} - \frac{dG_{s(b)}}{dG_0} \right| + \omega_l \left(\|\Upsilon(a-b)\|_\perp \right),$$

where (a) follows by replacing first the law $G_{s(a)}$ of $\mathcal{G}$ with $G_{s(b)}$ and then Υa with Υb, using $\left| \|T - \Upsilon a\|_\perp - \|T - \Upsilon b\|_\perp \right| \leq \|\Upsilon(a-b)\|_\perp$. The right hand side does not depend on T. Moreover, it is continuous in (a,b) and vanishes for $a = b$. For the first term, this follows from the expression of $dG_{s(a)}/dG_0$ in (B.12) below and the dominated convergence theorem. For the second term, it holds since $\|\Upsilon \cdot\|_\perp$ is a seminorm on the finite dimensional space $\mathbb{R}^{d_a}$. Hence, for any $r > 0$ and $\varepsilon > 0$, there exists a finite set $A \subset \{a : \|a\| \leq r\}$ such that, for any T,

$$\sup_{\|a\|\leq r} R(T,a) \leq \max_{a\in A} R(T,a) + \varepsilon.$$

Then, for any $\lambda > 0$,

$$\begin{aligned}\mathcal{R}(l) + \varepsilon &\overset{(b)}{\geq} \inf_T \sup_{\|a\| \leq r} R(T,a) \\ &\overset{(c)}{\geq} \inf_T \int R(T,a)\, dN(0, \lambda \mathrm{Id}_{d_a})(a) - \sup |l| \cdot N(0, \lambda \mathrm{Id}_{d_a}) \left(\{a : \|a\| > r\}\right) \\ &\overset{(d)}{\geq} \mathbb{E}\left[l \circ \|\Upsilon V_\lambda\|_\perp\right] - \sup |l| \cdot N(0, \lambda \mathrm{Id}_{d_a}) \left(\{a : \|a\| > r\}\right),\end{aligned}$$

where (b) follows from (B.10) and the choice of A, (c) from splitting the integral into $\{\|a\| \leq r\}$ and its complement, and (d) from Lemma B.7, with Id_{d_a} and V_λ as there. Let $r \to \infty$ and $\varepsilon \to 0$, and then $\lambda \to \infty$. Then V_λ converges in law to $Z \sim N(0, I_a^{-1})$, while $l \circ \|\Upsilon \cdot\|_\perp$ is bounded and continuous. This gives $\mathcal{R}(l) \geq \mathbb{E}[l \circ \|\Upsilon Z\|_\perp]$.

It remains to extend the bound to a general loss function. Let $l : [0, \infty) \mapsto [0, \infty)$ be non-decreasing and lower semicontinuous. For $k \in \mathbb{N}$, let

$$l_k(t) := \min\left\{k,\ \inf_{t' \geq 0} \left(l(t') + k|t - t'|\right)\right\}, \quad t \geq 0.$$

Each l_k is bounded, k-Lipschitz and non-decreasing, and $l_k \leq l_{k+1} \leq l$. Moreover, $l_k(t) \to l(t)$ for every $t \geq 0$. To see this, let $c < l(t)$. By the lower semicontinuity of l, there is $\delta > 0$ with $l(t') > c$ for $|t' - t| < \delta$. Hence $l(t') + k|t - t'| \geq \min\{c, k\delta\}$ for all $t' \geq 0$, so that $l_k(t) \geq c$ for $k \geq \max\{c, c/\delta\}$. By the bound above applied to l_k, $\mathcal{R}(l) \geq \mathcal{R}(l_k) \geq \mathbb{E}[l_k \circ \|\Upsilon Z\|_\perp]$ for every k. The monotone convergence theorem provides the result. □

B.3.2. *Auxiliary lemmas.* The proof of Theorem 4.6 reduces the asymptotic risk to a Bayes risk in the limit experiment, which is bounded below by the following two lemmas. They extend Anderson's lemma (Lemma 8.5 in van der Vaart (1998)) and the Bayesian argument in the proof of Proposition 8.6 there to estimators that need not lie in the range of Υ. The first lemma is a consequence of the log-concavity of the Gaussian density.

Lemma B.6. *Let $V \sim N(0, \Sigma)$ on $\mathbb{R}^d$ with Σ positive definite, let ν be a seminorm on $\mathbb{R}^d$ and $c \geq 0$, and let $B \subset \mathbb{R}^d$ be a non-empty convex set with $\nu(b_1 - b_2) \leq 2c$ for any $b_1, b_2 \in B$. Then $\mathbb{P}(V \in B) \leq \mathbb{P}(\nu(V) \leq c)$.*

Proof. In terms of sets, Anderson's lemma states that $\mathbb{P}(V \in h + K) \leq \mathbb{P}(V \in K)$ for every $h \in \mathbb{R}^d$ and every convex set K symmetric about 0. With $K := \{x : \nu(x) \leq c\}$, it gives the lemma when B lies in a translate of K. A convex set with ν-diameter at most $2c$ need not lie in such a translate[23]. I therefore compare B with the symmetric set $\frac{1}{2}B - \frac{1}{2}B \subset K$ instead, using the Prékopa–Leindler inequality.

Let $C := \{\frac{1}{2}b_1 - \frac{1}{2}b_2 : b_1, b_2 \in B\}$, which is convex, and hence Lebesgue measurable, and satisfies $C \subset \{x : \nu(x) \leq c\}$ by the assumption on B. Let ϕ be the density of V.

[23]For example, in the Euclidean norm on $\mathbb{R}^2$, an equilateral triangle with side $2c$ lies in no ball of radius c.

Since $\log\phi$ is concave, $\phi(\frac{1}{2}x+\frac{1}{2}y) \geq \phi(x)^{1/2}\phi(y)^{1/2}$ for any $x, y \in \mathbb{R}^d$, and $\frac{1}{2}x+\frac{1}{2}y \in C$ for $x \in B$ and $y \in -B$. The Prékopa–Leindler inequality (Theorem 3.15 in Wainwright (2019) with $\lambda = 1/2$) applied to $\phi\mathbb{1}_B$, $\phi\mathbb{1}_{-B}$ and $\phi\mathbb{1}_C$ then gives

$$\mathbb{P}(\nu(V) \leq c) \geq \mathbb{P}(V \in C) \geq \mathbb{P}(V \in B)^{1/2}\,\mathbb{P}(V \in -B)^{1/2} = \mathbb{P}(V \in B),$$

where the last equality follows from $V \stackrel{d}{=} -V$. □

Lemma B.7. *Suppose that Assumption 4.5 holds, let* $I_a := (P_0 s_j s_k)_{j,k=1}^{d_a}$ *and, for* $a \in \mathbb{R}^{d_a}$, *let* $s(a) := \sum_{j=1}^{d_a} a_j s_j$. *For a randomized statistic* $T = T(\mathcal{G}, U)$ *with values in* $\mathcal{C}(\Theta_I)$, *where* U *is uniformly distributed on* $[0,1]$ *and independent of* $\mathcal{G}$, *and a non-decreasing and lower semicontinuous loss function* $l : [0,\infty) \mapsto [0,\infty)$, *let*

$$R(T,a) := \mathbb{E}\left[l \circ \|T - \Upsilon a\|_\perp\right], \quad a \in \mathbb{R}^{d_a}, \tag{B.11}$$

where the expectation is taken under $\mathcal{G} \sim G_{s(a)}$ *in (16). Then for any such* T *and* l *and any* $\lambda > 0$, *it holds that*

$$\int R(T,a)\, dN(0, \lambda \mathrm{Id}_{d_a})(a) \geq \mathbb{E}\left[l \circ \|\Upsilon V_\lambda\|_\perp\right], \quad V_\lambda \sim N\left(0, (I_a + \lambda^{-1}\mathrm{Id}_{d_a})^{-1}\right),$$

where Id_{d_a} *is the* $d_a \times d_a$ *identity matrix.*

Proof. In the proof of Proposition 8.6 in van der Vaart (1998), the mean h of a $N(h,\Sigma)$ experiment is given a normal prior, and $T - Ah$ splits into the posterior noise and an independent remainder, so that Anderson's lemma bounds the Bayes risk. Here, the limit experiment is again a normal experiment, and the posterior of a is again normal. However, T takes values in $\mathcal{C}(\Theta_I)$ and need not lie in the range of Υ, so $T - \Upsilon a$ does not split in this way. Instead, I bound the posterior risk given the data by Lemma B.6, applied to the convex sets $\{b : \|T - \Upsilon b\|_\perp \leq c\}$.

I first express the limit experiment in terms of a normal vector. Let

$$Y := I_a^{-1}\left(\mathcal{G}(s_1), \cdots, \mathcal{G}(s_{d_a})\right)^t.$$

Since

$$\mathbb{E}\left(\mathcal{G}(s(a)) - \sum\nolimits_j a_j \mathcal{G}(s_j)\right)^2 = P_0\left(s(a) - \sum\nolimits_j a_j s_j\right)^2 = 0,$$

it holds that $\mathcal{G}(s(a)) = a^t I_a Y$ G_0-a.s., and hence, by (16),

$$\frac{dG_{s(a)}}{dG_0} = \exp\left(a^t I_a Y - \frac{1}{2}a^t I_a a\right), \quad a \in \mathbb{R}^{d_a}, \tag{B.12}$$

using $P_0 s(a)^2 = a^t I_a a$; in particular $Y \sim N(a, I_a^{-1})$ under $G_{s(a)}$.

I next compute the posterior. Let $A \sim N(0, \lambda\mathrm{Id}_{d_a})$, let $\mathcal{G} \sim G_{s(a)}$ conditionally on $A = a$, and let U be uniformly distributed on $[0,1]$ and independent of $(A, \mathcal{G})$. Then the left hand side of the inequality in Lemma B.7 equals $\mathbb{E}[l \circ \|T - \Upsilon A\|_\perp]$. By (B.12), the conditional density of A given $(\mathcal{G}, U)$ is proportional to $\exp(a^t I_a Y - \frac{1}{2}a^t(I_a + \lambda^{-1}\mathrm{Id}_{d_a})a)$, which depends on $(\mathcal{G}, U)$ only through Y; that is, $A - \mu_\lambda(Y)$ given $(\mathcal{G}, U)$ is distributed as V_λ, where $\mu_\lambda(Y) := (I_a + \lambda^{-1}\mathrm{Id}_{d_a})^{-1} I_a Y$.

Finally, I bound the posterior risk. Fix $(\mathcal{G}, U)$ and $c \geq 0$, and let $B_c := \{b \in \mathbb{R}^{d_a} : \|T - \Upsilon b\|_\perp \leq c\}$, which is convex since $b \mapsto \|T - \Upsilon b\|_\perp$ is convex. For $b_1, b_2 \in B_c$, the triangle inequality for $\|\cdot\|_\perp$ gives

$$\|\Upsilon(b_1 - b_2)\|_\perp \leq \|T - \Upsilon b_1\|_\perp + \|T - \Upsilon b_2\|_\perp \leq 2c.$$

If B_c is non-empty, Lemma B.6 applied to $B_c - \mu_\lambda(Y)$ with the seminorm $\nu := \|\Upsilon \cdot\|_\perp$ shows

$$\mathbb{P}\left(\|T - \Upsilon A\|_\perp \leq c \mid \mathcal{G}, U\right) = \mathbb{P}\left(A - \mu_\lambda(Y) \in B_c - \mu_\lambda(Y) \mid \mathcal{G}, U\right) \leq \mathbb{P}\left(\|\Upsilon V_\lambda\|_\perp \leq c\right),$$

which holds trivially if B_c is empty. Since this holds for any $c \geq 0$, $\|T - \Upsilon A\|_\perp$ given $(\mathcal{G}, U)$ is stochastically larger than $\|\Upsilon V_\lambda\|_\perp$; as l is non-decreasing and non-negative, this implies $\mathbb{E}[l \circ \|T - \Upsilon A\|_\perp \mid \mathcal{G}, U] \geq \mathbb{E}[l \circ \|\Upsilon V_\lambda\|_\perp]$. Integrating over $(\mathcal{G}, U)$ provides the result. □

B.4. **Proofs for Section 4.3.**

B.4.1. *Proofs for the main text.*

Proof. Lemma 4.10. For $\eta \in \mathbb{R}^{d_a}$ near 0, write

$$\Phi(\theta,\eta) := q\left(P_0 m(\theta,\cdot) + \sum\nolimits_{j=1}^{d_a} \eta_j P_0[m(\theta,\cdot)s_j]\right).$$

Then $\Phi(\theta,\eta) = \Phi(\theta, P_{1,\eta})$, $\Phi(\theta, ta) = \Phi(\theta, P_{t,a})$ and $\Phi(\theta,0) = \Phi(\theta,P_0)$. By the assumptions, there exist a compact neighborhood $N_0 \subset \Theta^o$ of Θ_I and a compact neighborhood V_0 of $0 \in \mathbb{R}^{d_a}$ such that $(\theta,\eta) \mapsto \Phi(\theta,\eta)$ is continuous on $\Theta \times V_0$ and C^1 on $N_0 \times V_0$. Moreover, $\nabla_\eta \Phi(\theta,0)a = \frac{d}{dt}\Phi(\theta, P_{t,a})|_{t=0}$. Let ω be a modulus of continuity of the derivative of Φ on $N_0 \times V_0$. For $\theta \in N_0$, let $G_j(\theta) := \nabla_\theta \Phi_j(\theta,0)$, and let $B(\theta) := \mathcal{J}_1 \cup \mathcal{A}(\theta, P_0)$ be the set of binding restrictions at $\theta \in \Theta_I$. Fix $a \in \mathbb{R}^{d_a}$, and write $\Theta_I(t) := \Theta_I(\Phi, P_{t,a})$.

The proof proceeds as follows. I first construct Υ so that, to the first order, every binding restriction stays binding when each point y of Θ_I moves to $y + t(\Upsilon a)(y)$ and the distribution moves to $P_{t,a}$; this is (a). I then show an error bound: near Θ_I, the distance from a point to the identified set is at most a constant times the violation of the restrictions at that point. By (a), the moved points violate the restrictions under $P_{t,a}$ only by $o(t)$, and the points of $\Theta_I(t)$, moved back, violate the restrictions under P_0 only by $o(t)$. The error bound then turns these two facts into the two directed Hausdorff distances in (b).

The construction of Υ starts from local versions. Fix $\vartheta \in \Theta_I$. By Assumption 4.9 (iii), the gradients $G_j(\vartheta)$, $j \in B(\vartheta)$, are linearly independent. Moreover, $\Phi_j(\vartheta,0) < 0$ for $j \in \mathcal{J}_2 \setminus \mathcal{A}(\vartheta, P_0)$. By the continuity of G_j and Φ_j, there exist an open ball $V_\vartheta \subset N_0$ around ϑ and $c_\vartheta > 0$ such that, for every $y \in \overline{V}_\vartheta$, the gradients $G_j(y)$, $j \in B(\vartheta)$, are linearly independent and $\Phi_j(y,0) \le -c_\vartheta$ for $j \in \mathcal{J}_2 \setminus \mathcal{A}(\vartheta, P_0)$. By compactness, Θ_I is covered by finitely many such balls $V_1, \cdots, V_L$ with centers $\vartheta_1, \cdots, \vartheta_L$. Let $B_l := B(\vartheta_l)$ and $c := \min_l c_{\vartheta_l} > 0$. If $y \in \Theta_I \cap V_l$, then $B(y) \subset B_l$, since $\Phi_j(y,0) \le -c < 0$ for $j \notin B_l$. For $y \in \overline{V}_l$, let $G_l(y)$ be the matrix with rows $G_j(y)^t$, $j \in B_l$, let $\Phi_{B_l} := (\Phi_j)_{j \in B_l}$, and define

$$g_l(y)[a] := -G_l(y)^t \left(G_l(y) G_l(y)^t\right)^{-1} \nabla_\eta \Phi_{B_l}(y,0)a,$$

with $g_l := 0$ if $B_l = \emptyset$. Then $G_l(y) g_l(y)[a] = -\nabla_\eta \Phi_{B_l}(y,0)a$. That is, moving y by $t g_l(y)[a]$ offsets, to the first order, the change of the restrictions in B_l caused by moving the distribution to $P_{t,a}$.

I next combine them. Let $\{\chi_l\}_{l=1}^L$ be a continuous partition of unity on Θ_I subordinate to $\{V_l\}_{l=1}^L$, and define $(\Upsilon a)(y) := \sum_l \chi_l(y) g_l(y)[a]$ for $y \in \Theta_I$. Then $\Upsilon a \in \mathcal{C}(\Theta_I)$, and $a \mapsto \Upsilon a$ is linear. Let $y \in \Theta_I$, and let j be any restriction. Then one of the following holds. (α) $j \in B_l$ for every l with $\chi_l(y) > 0$. Averaging $G_j(y)^t g_l(y)[a] = -\nabla_\eta \Phi_j(y,0)a$ over these l then gives

$$G_j(y)^t (\Upsilon a)(y) = -\nabla_\eta \Phi_j(y,0)a. \tag{B.13}$$

(β) $j \notin B_l$ for some l with $\chi_l(y) > 0$. Then $j \in \mathcal{J}_2 \setminus \mathcal{A}(\vartheta_l, P_0)$ and $y \in V_l$, so that $\Phi_j(y,0) \le -c$. Since $B(y) \subset B_l$ whenever $\chi_l(y) > 0$, every $j \in B(y)$ falls in case (α). This shows (a).

The key tool for (b) is the following error bound: there exist $\kappa < \infty$ and $\varepsilon > 0$ such that

$$d\left(z, \Theta_I(\Phi, P_{1,\eta})\right) \leq \kappa \left\| \left(\Phi_{\mathcal{J}_1}(z,\eta)^t, \Phi_{\mathcal{J}_2}(z,\eta)_+^t\right)^t \right\| \tag{B.14}$$

whenever $d(z, \Theta_I) \leq \varepsilon$, $\|\eta\| \leq \varepsilon$ and the right-hand side is at most $\kappa\varepsilon$. Here $(\cdot)_+$ is taken componentwise. The norm on the right-hand side measures the violation of the restrictions at (z, η). To prove the bound, fix $\vartheta \in \Theta_I$. Consider the constraint system $\Phi_{\mathcal{J}_1}(\theta, \eta) = 0$, $\Phi_{\mathcal{J}_2}(\theta, \eta) \leq 0$, $\eta = \eta'$ in the unknown (θ, η), with η' as a parameter. Its solution set is $\{\theta : \Phi_{\mathcal{J}_1}(\theta, \eta') = 0, \Phi_{\mathcal{J}_2}(\theta, \eta') \leq 0\} \times \{\eta'\}$. This system is of the form in Example 9.44 in Rockafellar and Wets (1998). Its constraint qualification holds at $(\vartheta, 0)$, as I now show. Let $(w_1^t, w_2^t)^t$ be a vector normal to the constraint set at $(\Phi(\vartheta, 0)^t, 0^t)^t$. Then $w_{1,j} = 0$ for $j \notin B(\vartheta)$. Suppose that this vector is annihilated by the transposed Jacobian of $(\theta, \eta) \mapsto (\Phi(\theta, \eta)^t, \eta^t)^t$ at $(\vartheta, 0)$. Then $\sum_{j \in B(\vartheta)} w_{1,j} G_j(\vartheta) = 0$, so that $w_1 = 0$ by Assumption 4.9 (iii). It follows that $w_2 = -\nabla_\eta \Phi(\vartheta, 0)^t w_1 = 0$. Hence (B.14) holds near $(\vartheta, 0)$ with some κ_ϑ and ε_ϑ. Metric regularity is a local property, so the system need only be C^1 near $(\vartheta, 0)$. A finite cover of Θ_I by balls of radius $\varepsilon_\vartheta/2$ then provides κ and ε. I take ε small enough that every point within $(1 + \kappa)\varepsilon$ of Θ_I lies in $N_0 \subset \Theta^o$. Then the nearest solution of the system lies in $\Theta_I(\Phi, P_{1,\eta})$.

With (a) and the error bound, the first directed distance follows. Let $C < \infty$ bound $\|((\Upsilon a)(y)^t, a^t)^t\|$ over $y \in \Theta_I$ and the norm of the derivative of Φ on $N_0 \times V_0$. Let $y \in \Theta_I$ and $z := y + t(\Upsilon a)(y)$. The increment from $(y, 0)$ to (z, ta) is $t((\Upsilon a)(y), a)$, of norm at most Ct. By the mean value theorem, for every j,

$$\left|\Phi_j(z, ta) - \Phi_j(y, 0) - t\left\{G_j(y)^t(\Upsilon a)(y) + \nabla_\eta \Phi_j(y, 0)a\right\}\right| \leq Ct \cdot \omega(Ct).$$

In case (α), the term in braces vanishes by (B.13). Moreover, $\Phi_j(y, 0) = 0$ for $j \in \mathcal{J}_1$ and $\Phi_j(y, 0) \leq 0$ for $j \in \mathcal{J}_2$. In case (β), $\Phi_j(z, ta) \leq -c + C^2 t < 0$ for small t. Hence the violation of the restrictions at (z, ta) is at most $\sqrt{d_q} Ct \cdot \omega(Ct)$. For small t, (B.14) therefore yields

$$\sup_{y \in \Theta_I} d\left(y + t(\Upsilon a)(y), \Theta_I(t)\right) \leq \kappa\sqrt{d_q} Ct \cdot \omega(Ct) = o(t). \tag{B.15}$$

The other direction is similar, with the roles of P_0 and $P_{t,a}$ exchanged. I first note that $\varepsilon_t' := \sup_{z \in \Theta_I(t)} d(z, \Theta_I) \to 0$. Otherwise, there are $t_k \searrow 0$ and $z_k \in \Theta_I(t_k)$ with $d(z_k, \Theta_I)$ bounded away from zero. Since Θ is compact, I may assume $z_k \to \bar{z} \in \Theta$. By continuity, $\bar{z}$ satisfies the restrictions under P_0, i.e., $\bar{z} \in \Theta_I$, a contradiction. Now let $z \in \Theta_I(t)$, let $y \in \Theta_I$ be a nearest point to z, so that $\|z - y\| \leq \varepsilon_t'$, and let $y' := z - t(\Upsilon a)(y)$. The increment from (z, ta) to $(y', 0)$ is $-t((\Upsilon a)(y), a)$, of norm at most Ct, and every point of the segment between them lies within $\varepsilon_t' + Ct$ of $(y, 0)$. By the mean value theorem along this segment, comparing the derivative of Φ with that at $(y, 0)$, for every j,

$$\left|\Phi_j(y', 0) - \Phi_j(z, ta) + t\left\{G_j(y)^t(\Upsilon a)(y) + \nabla_\eta \Phi_j(y, 0)a\right\}\right| \leq Ct \cdot \omega(\varepsilon_t' + Ct).$$

In case (α), the term in braces vanishes, while z satisfies the restrictions under $P_{t,a}$. In case (β), $\Phi_j(y', 0) \leq -c + C(\varepsilon_t' + Ct) < 0$ for small t. Hence the violation of the restrictions at $(y', 0)$ is at most $\sqrt{d_q} Ct \cdot \omega(\varepsilon_t' + Ct) = o(t)$. Since $d(y', \Theta_I) \leq \varepsilon_t' + Ct$,

(B.14) with $\eta = 0$ provides $y'' \in \Theta_I$ with $\|y' - y''\| = o(t)$, uniformly in z. Then $\|y - y''\| \le \varepsilon'_t + Ct + o(t)$, and

$$\|z - (y'' + t(\Upsilon a)(y''))\| \le \|y' - y''\| + t\,\|(\Upsilon a)(y) - (\Upsilon a)(y'')\|\,.$$

By the uniform continuity of Υa on Θ_I, it follows that

$$\sup_{z \in \Theta_I(t)} d\left(z, \Theta_I \oplus_t \Upsilon a\right) = o(t). \tag{B.16}$$

Combining (B.15) and (B.16) with $t = 1/\sqrt{n}$ provides (b). □

Proof. Proposition 4.11. I apply Lemma B.8 to the paths $t \mapsto P_{t,a}$ of Section 4.3 with $K_{t,a} := \Theta_I(\Phi, P_{t,a})$. The set Θ_I satisfies the condition in Lemma 4.3 by Assumption 4.9 (i). For $a \in \mathbb{R}^{d_a}$ and large n, $dP_{1/\sqrt{n},a}/dP_0 = 1 + n^{-1/2}\sum_j a_j s_j$, and $|\sqrt{1+u} - 1 - u/2| \le u^2$ for $|u| \le 1/2$; since $\sum_j a_j s_j$ is bounded, it follows that

$$\begin{aligned}&\sup_{x \in \mathcal{X}} \left| \sqrt{n}\left(\sqrt{(dP_{1/\sqrt{n},a}/dP_0)(x)} - 1\right) - \frac{1}{2}\sum\nolimits_j a_j s_j(x) \right| \\ &\quad \le n^{-1/2} \sup_{x \in \mathcal{X}} \left| \sum\nolimits_j a_j s_j(x) \right|^2 \to 0,\end{aligned}$$

which implies the condition (a). The set $K_{t,a}$ is closed in Θ, since $\theta \mapsto \Phi(\theta, P_{t,a})$ is continuous on Θ by the assumptions of Lemma 4.10, and it is non-empty for small t by (B.15) in the proof of Lemma 4.10, so that $K_{t,a} \in \mathcal{K}_\Theta$; the condition (b) is Lemma 4.10 (b). Hence Lemma B.8 provides the result. □

Proof. Theorem 4.12. Recall that $\mathcal{G}$ is the centered Gaussian process indexed by $\{s \in L^2(P_0) : P_0 s = 0\}$ with covariance $\mathbb{E}\,\mathcal{G}(s)\mathcal{G}(s') = P_0(ss')$, as defined in Section 4.1. Below, $\mathcal{G}$, $L^2(P_0)$-projections and expectations under P_0 act componentwise on $\mathbb{R}^{d_m}$-valued functions. For $\vartheta \in \Theta_I$, write $\tilde{m}_\vartheta := m(\vartheta, \cdot) - P_0 m(\vartheta, \cdot)$ and $B(\vartheta) := \mathcal{J}_1 \cup \mathcal{A}(\vartheta, P_0)$. By Assumption 4.9 (ii) and (v), $\vartheta \mapsto \tilde{m}_\vartheta$ is continuous from Θ_I into $L^2(P_0)$. Moreover, $\vartheta \mapsto \mathcal{G}(\tilde{m}_\vartheta)$ has a version with continuous paths on Θ_I; this version is used in (19) and throughout. Finally, $\mathcal{G}$ is linear almost surely on any finite set of functions. Indeed, by the covariance of $\mathcal{G}$, $\mathrm{Var}(\mathcal{G}(g + g') - \mathcal{G}(g) - \mathcal{G}(g')) = P_0((g + g') - g - g')^2 = 0$, and likewise $\mathrm{Var}(\mathcal{G}(rg) - r\mathcal{G}(g)) = 0$ for $r \in \mathbb{R}$.

The proof extends that of Theorems 25.20 and 25.21 in van der Vaart (1998) (Section 25.3 there), which apply the bounds for parametric models to submodels whose scores span all directions. Two features are new here. Since the target is a field over Θ_I, the projections onto the scores must converge uniformly in ϑ. Moreover, the comparison with V goes through the seminorm $\|\cdot\|_\perp$. The argument has four parts. Proposition 4.11 gives a bound along finitely many paths, in terms of the field ΥZ. I rewrite this field through $\mathcal{G}$: it solves (19) with $\tilde{m}_\vartheta$ replaced by its projection onto the span of the scores. Since $\|\cdot\|_\perp$ is determined by the binding components, $\|\Upsilon Z\|_\perp$ is then close to $\|V\|_\perp$ whenever this projection is close to $\tilde{m}_\vartheta$. Letting the scores span all directions makes the difference vanish.

The starting point is the bound along finitely many paths. Let $s_1, \cdots, s_{d_a} \in \mathcal{S}$ be linearly independent in $L^2(P_0)$, and let Υ be as in Lemma 4.10. Let $f_1, \cdots, f_k$ and $\epsilon_1, \cdots, \epsilon_k$ be the functions and constants defining $\mathcal{N}_{P_0}$. For $a \in \mathbb{R}^{d_a}$ and small $t > 0$, $P_{t,a}$ is a probability measure, and

$$\left| \int_{\mathcal{X}} f_i \left(dP_{t,a} - dP_0 \right) \right| = t \left| P_0 \left[f_i \sum\nolimits_j a_j s_j \right] \right| \\ \leq t \left\| f_i \right\|_{L^1(P_0)} \sup_{x \in \mathcal{X}} \left| \sum\nolimits_j a_j s_j(x) \right|, \quad i = 1, \cdots, k.$$

Hence, for each finite $D \subset \mathbb{R}^{d_a}$, there is n_D such that $P_{1/\sqrt{n},a} \in \mathcal{N}_{P_0}$ for $n \geq n_D$ and $a \in D$. For such n and a, also $\Theta_I(\Phi, P_{1/\sqrt{n},a}) \in \mathcal{K}_\Theta$, as in the proof of Proposition 4.11. The left-hand side of the inequality in Theorem 4.12 is therefore at least $\liminf_n \sup_{a \in D} \mathbb{E}_{P^n_{1/\sqrt{n},a}}[l \circ (\sqrt{n} \cdot d_H(T_n, \Theta_I(\Phi, P_{1/\sqrt{n},a})))]$ for every such D. Proposition 4.11 then gives

$$\liminf_{n \to \infty} \sup_{P \in \mathcal{N}_{P_0} : \Theta_I(\Phi, P) \in \mathcal{K}_\Theta} \mathbb{E}_{P^n} \left[l \circ \left(\sqrt{n} \cdot d_H \left(T_n, \Theta_I(\Phi, P) \right) \right) \right] \geq \mathbb{E} \left[l \circ \left\| \Upsilon Z \right\|_\perp \right], \tag{B.17}$$

where $Z \sim N(0, I_a^{-1})$ with $I_a = (P_0 s_j s_k)_{j,k=1}^{d_a}$.

I next express this bound in terms of $\mathcal{G}$. Since the right-hand side of (B.17) depends only on the law of Z, I take $Z := I_a^{-1}(\mathcal{G}(s_1), \cdots, \mathcal{G}(s_{d_a}))^t$. It is defined on the probability space of $\mathcal{G}$ and V, and it has the law $N(0, I_a^{-1})$. Let Π denote the orthogonal projection in $L^2(P_0)$ onto the span of $s_1, \cdots, s_{d_a}$. Then $\Pi g = \sum_j b_j s_j$ with $b := I_a^{-1}(P_0[s_1 g], \cdots, P_0[s_{d_a} g])^t$. By the linearity of $\mathcal{G}$, $\mathcal{G}(\Pi g) = b^t(\mathcal{G}(s_1), \cdots, \mathcal{G}(s_{d_a}))^t = \sum_j Z_j P_0[s_j g]$ almost surely. For $g = \tilde{m}_\vartheta$, I use the right-hand side, which is continuous in ϑ, as the version of $\mathcal{G}(\Pi \tilde{m}_\vartheta)$. By the chain rule,

$$\frac{d}{dt} \Phi_j(\vartheta, P_{t,a}) \Big|_{t=0} = \nabla q_j \left(P_0 m(\vartheta, \cdot) \right) \sum\nolimits_k a_k P_0[m(\vartheta, \cdot) s_k].$$

Moreover, $P_0[m(\vartheta, \cdot) s_k] = P_0[\tilde{m}_\vartheta s_k]$, as $P_0 s_k = 0$. Hence Lemma 4.10 (a) with $a = Z$ gives, almost surely, for every $\vartheta \in \Theta_I$ and $j \in B(\vartheta)$,

$$\nabla_\theta \Phi_j(\vartheta, P_0)^t (\Upsilon Z)(\vartheta) = -\nabla q_j \left(P_0 m(\vartheta, \cdot) \right) \mathcal{G} \left(\Pi \tilde{m}_\vartheta \right). \tag{B.18}$$

This is (19) with $\tilde{m}_\vartheta$ replaced by its projection $\Pi \tilde{m}_\vartheta$.

The comparison with V rests on the fact that $\| \cdot \|_\perp$ is determined by the binding components: there is $K < \infty$ such that, for every map $\Delta : \Theta_I \mapsto \mathbb{R}^{d_\theta}$,

$$\left\| \Delta \right\|_\perp \leq K \sup_{\vartheta \in \Theta_I} \max_{j \in B(\vartheta)} \left| \nabla_\theta \Phi_j(\vartheta, P_0)^t \Delta(\vartheta) \right|, \tag{B.19}$$

with $\| \cdot \|_\perp$ given by (17) and the maximum over the empty set equal to 0. Let $\vartheta \in \Theta_I$ and $u \in N(\vartheta)$ with $\|u\| \leq 1$. If $B(\vartheta) = \emptyset$, then $N(\vartheta) = \{0\}$, and there is nothing to show. Otherwise, let $G(\vartheta)$ be the matrix with rows $\nabla_\theta \Phi_j(\vartheta, P_0)^t$, $j \in B(\vartheta)$. By Rockafellar and Wets (1998, Theorem 6.14), $u = G(\vartheta)^t \lambda$ for some $\lambda \in \mathbb{R}^{|B(\vartheta)|}$. Hence

$$\left| u^t \Delta(\vartheta) \right| = \left| \lambda^t G(\vartheta) \Delta(\vartheta) \right| \leq \sqrt{d_q} \left\| \lambda \right\| \max_{j \in B(\vartheta)} \left| \nabla_\theta \Phi_j(\vartheta, P_0)^t \Delta(\vartheta) \right|.$$

It therefore suffices to bound $\|\lambda\|$ uniformly in ϑ. The matrix $G(\vartheta)$ has full row rank by Assumption 4.9 (iii). With $\sigma(\vartheta) > 0$ its smallest singular value, $\|\lambda\| \leq \|G(\vartheta)^t\lambda\|/\sigma(\vartheta) = \|u\|/\sigma(\vartheta) \leq 1/\sigma(\vartheta)$. To bound $\sigma(\vartheta)$ from below, take the balls $V_1, \cdots, V_L$ and the index sets $B_1, \cdots, B_L$ from the proof of Lemma 4.10. For $y \in \overline{V}_l$, the matrix with rows $\nabla_\theta \Phi_j(y, P_0)^t$, $j \in B_l$, has full row rank and is continuous in y. Its smallest singular value is therefore at least some $c_l > 0$ on the compact set $\overline{V}_l$. If $\vartheta \in V_l$, then $B(\vartheta) \subset B_l$, so $G(\vartheta)$ consists of some of the rows of this matrix at $y = \vartheta$. Removing rows does not decrease the smallest singular value of a matrix with full row rank, so $\sigma(\vartheta) \geq c_l$. This gives (B.19) with $K := \sqrt{d_q}/\min_l c_l$.

Since $\|\cdot\|_\perp$ satisfies the triangle inequality, (19), (B.18) and (B.19) give, almost surely,

$$\big|\|\Upsilon Z\|_\perp - \|V\|_\perp\big| \leq \|\Upsilon Z - V\|_\perp \leq K C_q \sup_{\vartheta \in \Theta_I} \|\mathcal{G}(\tilde{m}_\vartheta) - \mathcal{G}(\Pi \tilde{m}_\vartheta)\|\,, \tag{B.20}$$

where $C_q := \sup_{\vartheta \in \Theta_I} \|\nabla q(P_0 m(\vartheta, \cdot))\|$ is finite by Assumption 4.9 (ii) and (iv). Likewise, if V' is another random element of $\mathcal{C}(\Theta_I)$ satisfying (19), then $\|V - V'\|_\perp = 0$ by (B.19). Hence $\|V'\|_\perp = \|V\|_\perp$ almost surely, and the bound does not depend on the choice of V.

It remains to let the scores span all directions. Since $\mathcal{X}$ is a Borel subset of a Euclidean space, $L^2(P_0)$ is separable, and hence so is $\mathcal{S}$. Let $\mathcal{S}_0 \subset \mathcal{S}$ be countable and dense in $\mathcal{S}$. The closure of the span of $\mathcal{S}_0$ contains $\mathcal{S}$, so this span is dense in $\{s \in L^2(P_0) : P_0 s = 0\}$. Arranging $\mathcal{S}_0$ in a sequence and deleting each element that lies in the span of the preceding ones gives linearly independent $s_1, s_2, \cdots \in \mathcal{S}$ with the same span. Let Π_k denote the orthogonal projection onto the span of $s_1, \cdots, s_k$. Then $\Pi_k g \to g$ for every $g \in L^2(P_0)$ with $P_0 g = 0$. The convergence is uniform on $\{\tilde{m}_\vartheta : \vartheta \in \Theta_I\}$. Indeed, this set is compact in $L^2(P_0)$, as the continuous image of Θ_I. Moreover, $I - \Pi_k$ has norm at most 1, so that $\|(g - \Pi_k g) - (g' - \Pi_k g')\| \leq \|g - g'\|$. Covering the set by finitely many balls of radius $\eta > 0$ therefore gives

$$\sigma_k := \sup_{\vartheta \in \Theta_I} \|\tilde{m}_\vartheta - \Pi_k \tilde{m}_\vartheta\|_{L^2(P_0)} \to 0 \quad \text{as } k \to \infty. \tag{B.21}$$

For each k, let Υ_k and Z_k be Υ and Z above with the scores $s_1, \cdots, s_k$, and let $R_k(\vartheta) := \mathcal{G}(\tilde{m}_\vartheta) - \mathcal{G}(\Pi_k \tilde{m}_\vartheta)$. It equals $\mathcal{G}(\tilde{m}_\vartheta - \Pi_k \tilde{m}_\vartheta)$ almost surely for each ϑ, and it has continuous paths. Fix a component $i \in \{1, \cdots, d_m\}$, and write $R_{k,i}$ for the i-th component of R_k. It is a centered Gaussian process on Θ_I with standard deviation at most σ_k at each ϑ. Since $g \mapsto g - \Pi_k g$ and $g \mapsto g - P_0 g$ have norm at most 1, the standard deviation of $R_{k,i}(\vartheta) - R_{k,i}(\vartheta')$ is at most $\min\{2\sigma_k, \|m_i(\vartheta, \cdot) - m_i(\vartheta', \cdot)\|_{L^2(P_0)}\}$. I apply Dudley's entropy bound (van der Vaart and Wellner (2023, Corollary 2.2.9)) with the standard deviation semimetric of $R_{k,i}$. Its packing numbers at scale ε are at most $N_{[]}(\varepsilon/2, \mathcal{M}_i, L^2(P_0))$, and equal to 1 for $\varepsilon > 2\sigma_k$. For a fixed $\vartheta_0 \in \Theta_I$ and a universal constant K_0, this gives

$$\begin{aligned}
\mathbb{E} \sup_{\vartheta \in \Theta_I} |R_{k,i}(\vartheta)| &\leq \mathbb{E}|R_{k,i}(\vartheta_0)| + K_0 \int_0^{2\sigma_k} \sqrt{\log N_{[]}(\varepsilon/2, \mathcal{M}_i, L^2(P_0))}\, d\varepsilon \\
&\leq \sigma_k + 2K_0 \int_0^{\sigma_k} \sqrt{\log N_{[]}(\varepsilon, \mathcal{M}_i, L^2(P_0))}\, d\varepsilon.
\end{aligned}$$

By (B.21) and Assumption 4.9 (v), the right-hand side tends to 0. Hence

$$\mathbb{E} \sup_{\vartheta \in \Theta_I} \|R_k(\vartheta)\| \to 0,$$

and in particular $\sup_{\vartheta \in \Theta_I} \|R_k(\vartheta)\| \to 0$ in probability.

Finally, (B.20) with Υ_k, Z_k and Π_k gives

$$|\|\Upsilon_k Z_k\|_\perp - \|V\|_\perp| \le K C_q \sup_{\vartheta \in \Theta_I} \|R_k(\vartheta)\|.$$

Hence $\|\Upsilon_k Z_k\|_\perp \to \|V\|_\perp$ in probability, and hence in distribution. Since l is lower semicontinuous and non-negative, the portmanteau theorem (e.g. van der Vaart and Wellner (2023, Theorem 1.3.4)) gives $\liminf_k \mathbb{E}[l \circ \|\Upsilon_k Z_k\|_\perp] \ge \mathbb{E}[l \circ \|V\|_\perp]$. As (B.17) holds with $\Upsilon_k Z_k$ for every k, the result follows.

Existence of V. A field V as in Theorem 4.12 exists. In the construction of Υ in the proof of Lemma 4.10, replace $\nabla_\eta \Phi_{B_l}(y, 0)a$ by the vector $(\nabla q_j(P_0 m(y, \cdot)) \mathcal{G}(\tilde{m}_y))_{j \in B_l}$, which is continuous in $y \in \Theta_I \cap \overline{V}_l$. Then $V := \sum_l \chi_l g_l$ is a continuous function of the path of $\vartheta \mapsto \mathcal{G}(\tilde{m}_\vartheta)$, hence a random element of $\mathcal{C}(\Theta_I)$. It satisfies (19) by the argument for case (α) there. □

B.4.2. *Auxiliary lemmas.* The following lemma states Theorems 4.4 and 4.6 for a general family of local deviations; it is used in the proof of Proposition 4.11.

Lemma B.8. *Let $\Theta_I \in \mathcal{K}_\Theta$ satisfy the condition in Lemma 4.3, and let $s_1, \cdots, s_{d_a} \in L^2(P_0)$ be linearly independent. Suppose that there exists a linear map $\Upsilon : \mathbb{R}^{d_a} \mapsto \mathcal{C}(\Theta_I)$ such that, for every $a \in \mathbb{R}^{d_a}$, there exist probability measures $P_{t,a}$ on $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ with $P_{t,a} \ll P_0$ and sets $K_{t,a} \in \mathcal{K}_\Theta$, defined for small $t > 0$, satisfying*

(a). $\sqrt{n}\left(\sqrt{dP_{1/\sqrt{n},a}/dP_0} - 1\right) \to \frac{1}{2}\sum_{j=1}^{d_a} a_j s_j$ *in* $L^2(P_0)$ *as* $n \to \infty$,

(b). $\lim_{n\to\infty} \sqrt{n} \cdot d_H\left(K_{1/\sqrt{n},a}, \Theta_I \oplus_{1/\sqrt{n}} \Upsilon a\right) = 0$.

Then for any set-valued estimators $T_n : \mathcal{X}^n \mapsto \mathcal{K}_\Theta$ satisfying the condition (ii) in Theorem 4.4 and for any non-decreasing and lower semicontinuous loss function $l : [0, \infty) \mapsto [0, \infty)$, it holds that

$$\sup_{D \in \mathcal{D}} \liminf_{n\to\infty} \sup_{a \in D} \mathbb{E}_{P^n_{1/\sqrt{n},a}}\left[l \circ \left(\sqrt{n} \cdot d_H\left(T_n, K_{1/\sqrt{n},a}\right)\right)\right] \ge \mathbb{E}\left[l \circ \|\Upsilon Z\|_\perp\right],$$

where $Z \sim N(0, I_a^{-1})$ with $I_a := (P_0 s_j s_k)_{j,k=1}^{d_a}$, and $\mathcal{D} := \{D \subset \mathbb{R}^{d_a} : 0 < |D| < \infty\}$ is a directed set with the partial order $\subset$.

Proof. The result is Theorem 4.6 with the pairs (θ, h) replaced by $a \in \mathbb{R}^{d_a}$, the models $P_{\theta+h/\sqrt{n}}$ by $P_{1/\sqrt{n},a}$, the scores $\dot{\ell}_{\theta,h} h$ by $\sum_{j=1}^{d_a} a_j s_j$, the identified sets $\Theta_I(P_{\theta+h/\sqrt{n}})$ by $K_{1/\sqrt{n},a}$ and the fields $g_{\theta,h}$ by Υa, and the proofs of Theorems 4.4 and 4.6 apply with these replacements. Indeed, these proofs use Assumption 4.2 only through Lemma B.1, whose proof (with $\mu = P_0$) uses only the convergence in (a), through Lemma 4.3 applied to Θ_I, which holds by the condition on Θ_I, and through (14), which is (b); they use Assumption 4.5 only through Lemma B.7, whose proof uses only that the local deviation indexed by a has the score $\sum_{j=1}^{d_a} a_j s_j$ and the field Υa, which holds here by construction. □

B.5. **Proofs for Section 5.1.**

Proof. Proposition 5.3. The result follows from Theorems 3.1 and 3.2 in Chernozhukov et al. (2007), with $a_n = \sqrt{n}$. It therefore suffices to verify their Conditions C.1 and C.2. Their Condition C.3 is Assumption 5.2. Two ingredients are needed: a lower bound on $\|\Phi_v(\cdot, P)\|$ in terms of the distance to Θ_I, and a uniform bound on $\Phi(\cdot, \mathbb{P}_n) - \Phi(\cdot, P)$.

The first ingredient is the following bound: there exist constants $C, \delta > 0$ such that

$$\|\Phi_v(\theta, P)\| \geq C \cdot [d(\theta, \Theta_I) \wedge \delta] \quad \text{for any } \theta \in \Theta. \tag{B.22}$$

By Assumption 4.9 (ii) and (iv), $\Phi(\cdot, P)$ is continuous on Θ. It is also C^1 on an open neighborhood $U \subset \Theta^o$ of Θ_I; recall that $\Theta_I \subset \Theta^o$ by Assumption 4.9 (i). Let

$$D := \left\{ y \in \mathbb{R}^{d_q} : y_j = 0 \text{ for } j \in \mathcal{J}_1,\ y_j \leq 0 \text{ for } j \in \mathcal{J}_2 \right\}.$$

Then $d(\Phi(\theta, P), D) = \|\Phi_v(\theta, P)\|$ and $\Theta_I \cap U = \{\theta \in U : \Phi(\theta, P) \in D\}$. Fix $\vartheta \in \Theta_I$, and let y be a vector normal to D at $\Phi(\vartheta, P)$. Then $y_j = 0$ for $j \in \mathcal{J}_2 \setminus \mathcal{A}(\vartheta, P)$. Hence, if $\nabla_\theta \Phi(\vartheta, P)^t y = 0$, then

$$\sum\nolimits_{j \in \mathcal{J}_1 \cup \mathcal{A}(\vartheta, P)} y_j \nabla_\theta \Phi_j(\vartheta, P) = 0,$$

so that $y = 0$ by Assumption 4.9 (iii). This is the constraint qualification in Example 9.44 of Rockafellar and Wets (1998). That example therefore gives $\kappa_\vartheta < \infty$ and $\varepsilon_\vartheta > 0$ such that $d(\theta, \Theta_I) \leq \kappa_\vartheta \|\Phi_v(\theta, P)\|$ whenever $\|\theta - \vartheta\| \leq \varepsilon_\vartheta$. Here ε_ϑ is taken small enough that the relevant points lie in U. By compactness, Θ_I is covered by finitely many open balls $\{\theta : \|\theta - \vartheta_l\| < \varepsilon_{\vartheta_l}\}$, $l = 1, \cdots, L$. Let $\kappa := \max_l \kappa_{\vartheta_l}$, and let $\delta > 0$ be such that $\{\theta : d(\theta, \Theta_I) < \delta\}$ is contained in the union of these balls. Then $d(\theta, \Theta_I) \leq \kappa \|\Phi_v(\theta, P)\|$ whenever $d(\theta, \Theta_I) < \delta$. On the compact set $\{\theta \in \Theta : d(\theta, \Theta_I) \geq \delta\}$, the continuous function $\theta \mapsto \|\Phi_v(\theta, P)\|$ is positive. Hence it is bounded below by some $c_\delta > 0$. Thus (B.22) holds with $C := \min\{1/\kappa, c_\delta/\delta\}$. It continues to hold, with the same C, when δ is replaced by a smaller value.

I next control $\Phi(\cdot, \mathbb{P}_n)$ uniformly. By the P-Donsker property in Assumption 5.1 (ii),

$$\sup_{\theta \in \Theta} \|\mathbb{P}_n m(\theta, \cdot) - P m(\theta, \cdot)\| = O_P(1/\sqrt{n}).$$

Since q is continuous, it is uniformly continuous on a compact neighborhood of the compact set $\{Pm(\theta, \cdot) : \theta \in \Theta\}$. Hence

$$\sup_{\theta \in \Theta} \|\Phi(\theta, \mathbb{P}_n) - \Phi(\theta, P)\| = o_P(1). \tag{B.23}$$

Moreover, q is C^1 on a neighborhood of the compact set $A_0 := \{Pm(\theta, \cdot) : \theta \in \Theta_I\}$ by Assumption 4.9 (iv). So there exist $r > 0$ and $L < \infty$ such that $\|q(y) - q(y')\| \leq L\|y - y'\|$ whenever $d(y, A_0) \leq r$ and $\|y - y'\| \leq r$. This follows from the mean value theorem on the balls of radius $2r$ around the points of A_0. By the continuity of $\theta \mapsto Pm(\theta, \cdot)$, taking δ above smaller if necessary, $d(Pm(\theta, \cdot), A_0) \leq r$ whenever $d(\theta, \Theta_I) \leq \delta$. Therefore

$$\sup_{\theta \in \Theta:\, d(\theta, \Theta_I) \leq \delta} \sqrt{n}\, \|\Phi(\theta, \mathbb{P}_n) - \Phi(\theta, P)\| = O_P(1). \tag{B.24}$$

Since $|a_+ - b_+| \leq |a - b|$, (B.23) and (B.24) also hold with Φ_v in place of Φ.

It remains to verify Conditions C.1 and C.2. The bound (B.22) is the partial identification condition (4.5) of Chernozhukov et al. (2007), and (4.1) for equalities, with $\Phi_v(\theta, P)$ in place of $\mathbb{E}_P[m_i(\theta)]_+$. Given it, Conditions C.1 and C.2 follow as in Step 1 of the proof of Theorem 4.2 in Chernozhukov et al. (2007), with three replacements. First, the norm replaces its square, so that $a_n = \sqrt{n}$ and $\gamma = 1$. Second, Assumption 5.1 (iii) replaces their Condition M.2 (e). Third, the bounds (B.23) and (B.24) replace the P-Donsker bound on $\mathbb{G}_n[m_i(\theta)]$. One point needs care: (B.24) holds only on $\{\theta : d(\theta, \Theta_I) \leq \delta\}$. On $\{\theta : d(\theta, \Theta_I) > \delta\}$, the bound (B.23) suffices, since $\|\Phi_v(\cdot, P)\| \geq C\delta$ there by (B.22).

Hence Theorem 3.1 in Chernozhukov et al. (2007) gives the first result. The second result follows from Theorem 3.2 there, by adding Assumption 5.2 (Condition C.3) and noting that

$$\tau = \sqrt{n} \cdot c_n = \inf_{\theta \in \Theta} \sqrt{n} \cdot Q_n(\theta) \leq \sup_{\theta \in \Theta_I} \sqrt{n} \cdot Q_n(\theta) = O_P(1),$$

where the last bound is Condition C.1 (e). □

B.6. **Proofs for Section 5.2.**

Proof. Theorem 5.6.
For any $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ recall the stochastic field

$$Z_n(\theta) := \sqrt{n} \cdot \left(d(\theta, \widehat{\Theta}_I) - d(\theta, \Theta_I) \right)$$

defined in (25). By the definition of d_H, it follows that $\sup_{\theta \in \mathbb{R}^{d_\theta}} |Z_n(\theta)| = \sqrt{n} \cdot d_H(\widehat{\Theta}_I, \Theta_I)$. Lemma B.11 further implies that the domain of the supremum can be restricted to the $\sqrt{n}$-local neighborhood of the boundary $\partial\Theta_I$, $\mathcal{T}_n$ in (B.25). Thus with probability approaching 1 it holds that

$$\sqrt{n} \cdot \left(d_H(\widehat{\Theta}_I, \Theta_I) \right) = \sup_{x \in \mathcal{T}_n(M)} |Z_n(x)| = \sup_{(\theta,\psi) \in \partial\Theta_I \times \overline{B_M(0)}} |Z_n(\theta, \psi)| ,$$

where the second equality follows from Lemma B.13. This corresponds to the "generalized Chernoff regularity" (Condition S.1 in Chernozhukov et al. (2007)) in the interior parameter case (recall $\Theta_I \subset \Theta^o$).

The stochastic field $Z_n(\theta, \psi)$ is not equicontinuous with respect to their arguments; hence typical argument of uniform weak convergence on a function space $\mathcal{C}(\partial\Theta_I \times \overline{B_M(0)})$ or $\ell^\infty(\partial\Theta_I \times \overline{B_M(0)})$ does not work. Following Chernozhukov et al. (2007), I instead decompose $\partial\Theta_I$ by the set of binding restrictions; the limit is then a Lipschitz functional of D, so that the continuous mapping theorem applies and no finite dimensional approximation is needed. The key intermediate result is the following.

Proposition B.9. *Suppose that the same conditions as in Theorem 5.6 hold. Then for large enough $M > 0$ it holds that*

$$\sqrt{n} \cdot d_H(\widehat{\Theta}_I, \Theta_I) = \sup_{(\theta,\psi) \in \partial\Theta_I \times \mathbb{R}^{d_\theta}} |Z_n(\theta, \psi)| + o(1) \rightsquigarrow \sup_{(\theta,\psi) \in \partial\Theta_I \times \mathbb{R}^{d_\theta}} |Z(\theta, \psi)| ,$$

as $n \to \infty$, where $Z(\theta, \psi) = d(\psi, \Psi(\theta)) - d(\psi, T(\theta))$ as defined in Lemma B.14.

Proof. See Appendix B.6.2. □

This characterizes the limiting distribution of $\sqrt{n} \cdot d_H(\widehat{\Theta}_I, \Theta_I)$. Lemma B.15 further establishes the equivalence

$$\sup_{(\theta,\psi) \in \partial\Theta_I \times \mathbb{R}^{d_\theta}} |Z(\theta, \psi)| = \sup_{\theta \in \partial\Theta_I} V(\theta) = \sup_{\theta \in \partial\Theta_I} d_H(\Psi(\theta), T(\theta)),$$

where $\Psi(\theta)$ and $T(\theta)$ are the local limit of $\sqrt{n} \cdot (\widehat{\Theta}_I - \theta)$ and $\sqrt{n} \cdot (\Theta_I - \theta)$ at $\theta \in \partial\Theta_I$ as in Lemma B.12. Finally the local quantity $d_H(\Psi(\theta), T(\theta))$ at $\theta \in \partial\Theta_I$ is expressed via the supremum of the inner product of the unit normal vector and the combination of the Gaussian approximation error, i.e.

$$d_H(\Psi(\theta), T(\theta)) = \sup \left\{ \left| \alpha^t D_{\mathcal{A}(\theta)}(\theta) \right| : \alpha \geq 0, \ \left\| \alpha^t G_{\mathcal{A}(\theta)}(\theta) \right\| = 1 \right\} .$$

by Lemma B.15. Taking the supremum over $\theta \in \partial\Theta_I$ then shows the result. □

B.6.1. *Auxiliary lemmas.* Throughout this subsection, I assume $\mathcal{J}_1 = \emptyset$ as well as the same conditions in Theorem 5.6.

Lemma B.10. *For $\widehat{\Theta}_I$ and c_n, the following results hold:*

(i). *with probability approaching* 1, $c_n = \inf_{\theta' \in \Theta} Q_n(\theta') = 0$*; hence* $\sqrt{n} \cdot c_n \to 0$ *in probability,*

(ii). $\widehat{\Theta}_I(\tau)$ *with* $\tau = \sqrt{n} \cdot c_n$ *satisfies*

$$\widehat{\Theta}_I(\tau) = \left\{\theta \in \Theta \subset \mathbb{R}^{d_\theta} : Q_n(\theta) \leq c_n\right\} = \left\{\theta \in \Theta \subset \mathbb{R}^{d_\theta} : \max_{j \in \mathcal{J}_2} \Phi_j(\theta, \mathbb{P}_n) \leq 0\right\}.$$

Proof. For (i), note that for any $\theta \in \partial\Theta_I$ it holds that $d(0, \mathrm{conv}\{\phi_j(\theta) : j \in \mathcal{A}(\theta)\}) \geq c > 0$ by Assumption 5.4, implying that $0 \notin \mathrm{conv}\{\phi_j(\theta) : j \in \mathcal{A}(\theta)\}$. Now by Gordan's theorem (e.g., Borwein and Lewis (2006) Theorem 2.2.1), one of the following cases holds: (a) there exists $\psi \in \mathbb{R}^{d_\theta}$ such that $\langle \phi_j(\theta), \psi \rangle < 0$ for any $j \in \mathcal{A}(\theta)$, or (b) $0 \in \mathrm{conv}\{\phi_j(\theta) : j \in \mathcal{A}(\theta)\}$; but (b) does not hold, as shown above. For this ψ in (a), consider the Taylor expansion around $\theta \in \partial\Theta_I$,

$$\Phi_j(\theta + t \cdot \psi, P) = \Phi_j(\theta, P) + t \cdot \langle \phi_j(\theta), \psi \rangle + o(t).$$

For $j \notin \mathcal{A}(\theta)$ it holds that $\Phi_j(\theta, P) < 0$; hence $t > 0$ can be taken small enough to satisfy $\Phi_j(\theta + t \cdot \psi, P) < 0$. For $j \in \mathcal{A}(\theta)$, it holds that $\Phi_j(\theta, P) = 0$ and $\langle \phi_j(\theta), \psi \rangle < 0$; hence $\Phi_j(\theta + t \cdot \psi, P) < 0$. Let $\theta^* = \theta + t \cdot \psi$ be such that $\Phi_j(\theta^*, P) < 0$ for any $j \in \mathcal{J}_2$. By continuity in t it holds that $\theta^* \in \Theta_I^o$.

For any such $\theta^* \in \Theta_I^o$ let $\delta := -\max_{j \in \mathcal{J}_2} \Phi_j(\theta^*, P)$; then $\delta > 0$ due to the slackness shown above. Note here that $\|\Phi(\theta^*, \mathbb{P}_n) - \Phi(\theta^*, P)\| = O_P(1/\sqrt{n})$ by Assumption 5.1 (ii) and the delta method (Assumption 4.9 (iv)). Thus, with probability approaching 1 it holds that

$$\begin{aligned}\max_{j \in \mathcal{J}_2} \Phi_j(\theta^*, \mathbb{P}_n) &\leq \max_{j \in \mathcal{J}_2} \Phi_j(\theta^*, P) + \left[\max_{j \in \mathcal{J}_2} \Phi_j(\theta^*, \mathbb{P}_n) - \max_{j \in \mathcal{J}_2} \Phi_j(\theta^*, P)\right] \\ &\leq -\delta + O_P(1/\sqrt{n}) < 0.\end{aligned}$$

This implies $\Phi_v(\theta^*, \mathbb{P}_n) = 0$ and $Q_n(\theta^*) = 0$, which shows the result (i).

For (ii), note that the result (i) implies that with probability approaching 1

$$\begin{aligned}\widehat{\Theta}_I(\tau) = \left\{\theta \in \Theta \subset \mathbb{R}^{d_\theta} : Q_n(\theta) \leq c_n\right\} &= \left\{\theta \in \Theta \subset \mathbb{R}^{d_\theta} : Q_n(\theta) = 0\right\} \\ &= \left\{\theta \in \Theta \subset \mathbb{R}^{d_\theta} : \Phi_v(\theta, \mathbb{P}_n) = 0\right\},\end{aligned}$$

where the second equality follows from $\mathcal{J}_1 = \emptyset$ and $W = W_n + o_P(1)$ being positive definite. The construction of Φ_v shows that the last set of the above is equivalent to the set in the result in (ii). □

Lemma B.11. *For $M > 0$ let*

$$\mathcal{T}_n(M) := \left\{\theta \in \mathbb{R}^{d_\theta} : d(\theta, \partial\Theta_I) \leq M/\sqrt{n}\right\}. \tag{B.25}$$

Under the condition that $d_H(\widehat{\Theta}_I, \Theta_I) = O_P(1/\sqrt{n})$*, one can take* M *large enough such that*

$$d_H(\widehat{\Theta}_I, \Theta_I) = \sup_{\theta \in \mathcal{T}_n(M)} \left| d(\theta, \widehat{\Theta}_I) - d(\theta, \Theta_I) \right|,$$

with probability approaching 1*.*

Proof. By the definition of d_H it holds that $d_H(\widehat{\Theta}_I, \Theta_I) = \sup_{\theta \in \mathbb{R}^{d_\theta}} \left| d(\theta, \widehat{\Theta}_I) - d(\theta, \Theta_I) \right|$, which implies the left hand side is bigger than the right one.

For the other direction, if $d_H(\widehat{\Theta}_I, \Theta_I) = 0$ with probability 1 the result is trivial; hence assume $d_H(\widehat{\Theta}_I, \Theta_I) > 0$ at least with some probability $\epsilon(n) > 0$ with $\epsilon(n) = O(1/\sqrt{n})$. First note that

$$P\left\{ \Theta_I^{-\delta_n} \subset \widehat{\Theta}_I \right\} \to 1,$$

where $\delta_n = M/\sqrt{n}$ and $\Theta_I^{-\delta_n}$ denotes the δ_n-contraction; $\Theta_I^{-\delta_n} = \{\theta \in \Theta_I : d(\theta, \partial\Theta_I) \geq \delta_n\}$. To see this, note that for $\theta \in \Theta_I^{-\delta_n}$ Assumption 5.4 shows $\max_{j \in \mathcal{J}_2} \Phi_j(\theta, P) \leq -c \cdot \delta_n$. Since $\sup_{\theta \in \Theta} \|\Phi_n(\theta, \mathbb{P}_n) - \Phi(\theta, P)\| = O_P(1/\sqrt{n})$, it holds that

$$\begin{aligned} \max_{j \in \mathcal{J}_2} \Phi_j(\theta, \mathbb{P}_n) &\leq \max_{j \in \mathcal{J}_2} \Phi_j(\theta, P) + \left[\max_{j \in \mathcal{J}_2} \Phi_j(\theta, \mathbb{P}_n) - \max_{j \in \mathcal{J}_2} \Phi_j(\theta, P) \right] \\ &\leq -c \cdot \delta_n + O_P(1/\sqrt{n}) < 0 \end{aligned}$$

with probability approaching 1 for large M. Hence $\theta \in \widehat{\Theta}_I$ by Lemma B.10 and the result follows.

Now recall that the definition of d_H reads

$$d_H(\widehat{\Theta}_I, \Theta_I) = \max \left\{ \sup_{x \in \widehat{\Theta}_I} d(x, \Theta_I), \sup_{y \in \Theta_I} d(y, \widehat{\Theta}_I) \right\}.$$

Accordingly, two cases arise: (i) $d_H(\widehat{\Theta}_I, \Theta_I) = \sup_{x \in \widehat{\Theta}_I} d(x, \Theta_I)$ and (ii) $d_H(\widehat{\Theta}_I, \Theta_I) = \sup_{y \in \Theta_I} d(y, \widehat{\Theta}_I)$.

Suppose that the case (i) holds. Take $x^* \in \widehat{\Theta}_I$ such that $d(x^*, \Theta_I) = d_H(\widehat{\Theta}_I, \Theta_I)$. The restriction $d_H(\widehat{\Theta}_I, \Theta_I) > 0$ implies $x^* \notin \Theta_I$. Since

$$d(x^*, \partial\Theta_I) \leq d(x^*, \Theta_I) = d_H(\widehat{\Theta}_I, \Theta_I) \leq M/\sqrt{n},$$

it holds that $x^* \in \mathcal{T}_n$. Clearly $\left| d(x^*, \widehat{\Theta}_I) - d(x^*, \Theta_I) \right| = d_H(\widehat{\Theta}_I, \Theta_I)$ as $x^* \in \widehat{\Theta}_I$ and the desired result holds.

Suppose now that the case (ii) holds. Take $y^* \in \Theta_I$ such that $d(y^*, \widehat{\Theta}_I) = d_H(\widehat{\Theta}_I, \Theta_I)$. The restriction $d_H(\widehat{\Theta}_I, \Theta_I) > 0$ implies $y^* \notin \widehat{\Theta}_I$. Since $P\{\Theta_I^{-\delta_n} \subset \widehat{\Theta}_I\} \to 1$, it holds that $y^* \notin \Theta_I^{-\delta_n}$ for large $n \in \mathbb{N}$; but then $d(y^*, \partial\Theta_I) < \delta_n = M/\sqrt{n}$, implying $y^* \in \mathcal{T}_n$. Clearly $|d(y^*, \widehat{\Theta}_I) - d(y^*, \Theta_I)| = d_H(\widehat{\Theta}_I, \Theta_I)$ as $y^* \in \Theta_I$ and the desired result holds as well. □

Lemma B.12. *Suppose that Assumption 5.1 holds and that $\Phi(\theta, P)$ is fully differentiable at any $\theta \in \partial\Theta_I$. Then for any $M > 0$, uniformly for $\psi \in \overline{B_M(0)}$ it holds that*

$$\sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, \mathbb{P}_n) \rightsquigarrow \begin{cases} \langle \phi_j(\theta), \psi \rangle + D_j(\theta) & j \in \mathcal{A}(\theta), \\ -\infty & j \notin \mathcal{A}(\theta). \end{cases}$$

Hence with $c_n = 0$ with probability approaching 1 it holds that

$$\begin{aligned} \Psi_n(\theta) \cap \overline{B_M(0)} &:= \sqrt{n} \cdot (\widehat{\Theta}_I - \theta) \cap \overline{B_M(0)} \\ &\rightsquigarrow \Psi(\theta) \cap \overline{B_M(0)} = \left\{ \psi \in \overline{B_M(0)} : \langle \phi_j(\theta), \psi \rangle \leq -D_j(\theta),\ j \in \mathcal{A}(\theta) \right\} \\ &= \left\{ \psi \in \overline{B_M(0)} : G_{\mathcal{A}(\theta)}(\theta)\psi \leq -D_{\mathcal{A}(\theta)}(\theta) \right\}, \\ T_n(\theta) \cap \overline{B_M(0)} &:= \sqrt{n} \cdot (\Theta_I - \theta) \cap \overline{B_M(0)} \\ &\to T(\theta) \cap \overline{B_M(0)} = \{ \psi \in \overline{B_M(0)} : \langle \phi_j(\theta), \psi \rangle \leq 0,\ j \in \mathcal{A}(\theta) \} \\ &= \{ \psi \in \overline{B_M(0)} : G_{\mathcal{A}(\theta)}(\theta)\psi \leq 0 \}. \end{aligned} \tag{B.26}$$

Proof. Fix $\theta \in \partial\Theta_I$ and $M > 0$. Note that for any $\psi \in \overline{B_M(0)}$ it holds that

$$\begin{aligned} &\sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, \mathbb{P}_n) \\ &= \underbrace{\sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, P)}_{(a)} + \underbrace{\left[\sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, \mathbb{P}_n) - \sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, P) \right]}_{(b)}, \end{aligned}$$

for any $j \in \mathcal{J}_2$. For (a) it holds that

$$(a) = \sqrt{n} \cdot \Phi_j(\theta, P) + \langle \phi_j(\theta), \psi \rangle + \sqrt{n} \cdot o(\|\psi/\sqrt{n}\|),$$

while for (b) one obtains

$$\begin{aligned} (b) &= \sqrt{n} \cdot [\Phi_j(\theta, \mathbb{P}_n) - \Phi_j(\theta, P)] \\ &\quad + \sqrt{n} \cdot \left\{ \left[\Phi_j(\theta + \psi/\sqrt{n}, \mathbb{P}_n) - \Phi_j(\theta + \psi/\sqrt{n}, P) \right] - [\Phi_j(\theta, \mathbb{P}_n) - \Phi_j(\theta, P)] \right\} \\ &= D_{n,j}(\theta) + \sqrt{n} \cdot o_P(\|\psi/\sqrt{n}\|), \end{aligned}$$

where the last equality follows from the stochastic equicontinuity of the process $\Phi_j(\theta, \mathbb{P}_n) - \Phi(\theta, P)$ on Θ due to the Donsker property of m and smooth Φ. where $D_{n,j}(\theta) := \sqrt{n} \cdot [\Phi_j(\theta, \mathbb{P}_n) - \Phi_j(\theta, P)]$. Combining the results and noting that $D_{n,j}(\theta) \rightsquigarrow D_j(\theta)$ by Assumptions 4.9 (iv) and 5.1 (ii) and the delta method, read on a Skorokhod representation as in the proof of Lemma B.21, one obtains

$$\sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, \mathbb{P}_n) = \sqrt{n} \cdot \Phi_j(\theta, P) + \langle \phi_j(\theta), \psi \rangle + D_j(\theta) + o_P(1),$$

which holds uniformly for $\psi \in \overline{B_M(0)}$. If $j \in \mathcal{A}(\theta)$, $\Phi_j(\theta, P) = 0$ and the first term vanishes; if $j \notin \mathcal{A}$, then $\Phi_j(\theta, P) < 0$ and $\sqrt{n} \cdot \Phi_j(\theta, P) \to -\infty$. This gives the first result.

For the second result, recall that Lemma B.10 shows

$$\widehat{\Theta}_I = \left\{ \theta \in \mathbb{R}^{d_\theta} : \max_{j \in \mathcal{J}_2} \Phi_j(\theta, \mathbb{P}_n) \leq 0 \right\}.$$

From the construction $\Psi_n(\theta) = \sqrt{n} \cdot (\widehat{\Theta}_I - \theta)$ for $\theta \in \partial\Theta_I$, it follows that $\psi \in \Psi_n(\theta) \Leftrightarrow \theta + \psi/\sqrt{n} \in \widehat{\Theta}_I$. Hence

$$\begin{aligned} \Psi_n(\theta) &= \sqrt{n} \cdot (\widehat{\Theta}_I - \theta) \\ &= \left\{\psi \in \mathbb{R}^{d_\theta} : \sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, \mathbb{P}_n) \leq 0, \text{ for all } j \in \mathcal{J}_2\right\}. \end{aligned} \tag{B.27}$$

Now let $\psi_n \in \Psi_n(\theta) \cap \overline{B_M(0)}$ with $\psi_n \to \psi$. For $j \in \mathcal{A}(\theta)$, the first result combined with the uniform expansion of $\Phi_j(\theta + \psi/\sqrt{n}, \mathbb{P}_n)$ on $\overline{B_M(0)}$ above shows

$$0 \geq \sqrt{n} \cdot \Phi(\theta + \psi_n/\sqrt{n}, \mathbb{P}_n) \to \langle \phi_j(\theta), \psi \rangle + D_j(\theta).$$

Hence $\langle \phi_j(\theta), \psi \rangle \leq -D_j(\theta)$ for any $j \in \mathcal{A}(\theta)$, i.e. $\psi \in \Psi(\theta) \cap \overline{B_M(0)}$; this implies $\lim_{n\to\infty} \sqrt{n} \cdot (\widehat{\Theta}_I - \theta) \subset \Psi(\theta) \cap \overline{B_M(0)}$.

Next let $\psi \in \Psi(\theta) \cap \overline{B_M(0)}$. If $\langle \phi_j(\theta), \psi \rangle < -D_j(\theta)$ for any $j \in \mathcal{A}(\theta)$, then $\sqrt{n} \cdot \Phi(\theta + \psi/\sqrt{n}, \mathbb{P}_n) < 0$ for any $j \in \mathcal{J}_2$ and any large $n \in \mathbb{N}$; thus $\psi \in \Psi_n(\theta) \cap \overline{B_M(0)}$ by (B.27). The other case is $\psi \in \partial\Psi(\theta) \cap \overline{B_M(0)}$. As seen in the proof of Lemma B.10, Assumption 5.4 combined with the Gordan's theorem (e.g. Borwein and Lewis (2006) Theorem 2.2.1) implies that there exists $\psi \in \mathbb{R}^{d_\theta}$ such that $\langle \phi_j(\theta), \psi \rangle < 0$ for any $j \in \mathcal{A}(\theta)$, i.e. the tangent cone $T(\theta)$ has non-empty interior. Moreover Assumption 5.4 (iii) implies $G_{\mathcal{A}}(\theta)v = -D_{\mathcal{A}}(\theta)$ has a solution v. Therefore, $\Psi(\theta) = \{\psi : G_{\mathcal{A}}(\theta)\psi \leq -D_{\mathcal{A}}(\theta)\} = \{\psi : G_{\mathcal{A}}(\theta)(\psi - v) \leq 0\}$, which is a translation of $T(\theta)$. Hence take $\psi_n \in \Psi(\theta)^o \cap \overline{B_M(0)}$ such that $\psi_n \to \psi \in \partial\Psi(\theta) \cap \overline{B_M(0)}$ and that $\psi_n \in \Psi_n(\theta) \cap \overline{B_M(0)}$ for all large $n \in \mathbb{N}$; this construction gives $\psi \in \liminf_{n\to\infty} \Psi_n(\theta) \cap \overline{B_M(0)}$. Hence it has been shown $\Psi(\theta) \cap \overline{B_M(0)} \subset \lim_{n\to\infty} \sqrt{n} \cdot (\widehat{\Theta}_I - \theta) \cap \overline{B_M(0)}$.

The result on $T_n(\theta)$ and $T(\theta)$ follows in a similar manner. In particular, note that

$$\sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, P) = \sqrt{n} \cdot \Phi_j(\theta, P) + \langle \phi_j(\theta), \psi \rangle + o(1),$$

which holds uniformly for $\psi \in \overline{B_M(0)}$. Note also that, from the characterization $\Theta_I = \{\theta \in \mathbb{R}^{d_\theta} : \max_{j\in\mathcal{J}_2} \Phi_j(\theta, P) \leq 0\}$ and $T_n(\theta) = \sqrt{n} \cdot (\Theta_I - \theta)$ for $\theta \in \partial\Theta_I$, it follows that $\psi \in T_n(\theta) \Leftrightarrow \theta + \psi/\sqrt{n} \in \Theta_I$. Hence

$$\begin{aligned} T_n(\theta) &= \sqrt{n} \cdot (\Theta_I - \theta) \\ &= \left\{\psi \in \mathbb{R}^{d_\theta} : \sqrt{n} \cdot \Phi_j(\theta + \psi/\sqrt{n}, P) \leq 0, \text{ for all } j \in \mathcal{J}_2\right\}. \end{aligned}$$

The rest of the argument follows the same argument, where $\{\Psi_n, \Psi, D_{\mathcal{A}}, v\}$ is replaced by $\{T_n, T, 0 \in \mathbb{R}^{d_\theta}, 0 \in \mathbb{R}^{d_\theta}\}$. □

Lemma B.13. *Recall* $Z_n(x) := \sqrt{n} \cdot \left(d(x, \widehat{\Theta}_I) - d(x, \Theta_I)\right)$ *defined in (25). For the fixed compact local graph* $\partial\Theta_I \times \overline{B_M(0)}$, *define the local process of* Z_n *as*

$$Z_n(\theta, \psi) := Z_n(\theta + \psi/\sqrt{n}), \tag{B.28}$$

where $\Psi(\theta)$ *and* $T(\theta)$ *are as in Lemma B.12. Then it holds that* $Z_n(\theta, \psi) = d(\psi, \Psi_n(\theta)) - d(\psi, T_n(\theta))$, *where* $\Psi_n(\theta)$ *and* $T_n(\theta)$ *are as in Lemma B.12. Moreover,*

$$\sup_{x\in\mathcal{T}_n(M)} |Z_n(x)| = \sup_{(\theta,\psi)\in\partial\Theta_I\times\overline{B_M(0)}} |Z_n(\theta, \psi)|.$$

Proof. First, for $x \in \mathcal{T}_n(M)$, let $\theta = \theta(x) \in \partial\Theta_I$ such that $d(x, \partial\Theta_I) = d(x, \theta)$, and define $\psi = \sqrt{n}(x - \theta)$; clearly $x = \theta + \psi/\sqrt{n}$. This map $x \mapsto (\theta, \psi)$ is bijective from $\mathcal{T}_n(M)$ onto $\partial\Theta_I \times \overline{B_M(0)}$. Next, note also that since for any set A it holds that $d(x, A) = (1/\sqrt{n}) \cdot d(\psi, \sqrt{n}(A - \theta))$, one obtains

$$\begin{aligned} Z_n(x) &= Z_n(\theta + \psi/\sqrt{n}) \\ &= \sqrt{n} \cdot \left(d(\theta + \psi/\sqrt{n}, \widehat{\Theta}_I) - d(\theta + \psi/\sqrt{n}, \Theta_I) \right) \\ &= d(\psi, \Psi_n(\theta)) - d(\psi, T_n(\theta)). \end{aligned}$$

The first result then follows from $Z_n(\theta, \psi) := Z_n(\theta + \psi/\sqrt{n})$ in (B.28). Combining the arguments, one obtains

$$\begin{aligned} \sup_{x \in \mathcal{T}_n(M)} |Z_n(x)| &= \sup_{(\theta,\psi) \in \partial\Theta_I \times \overline{B_M(0)}} |d(\psi, \Psi_n(\theta)) - d(\psi, T_n(\theta))| \\ &= \sup_{(\theta,\psi) \in \partial\Theta_I \times \overline{B_M(0)}} |Z_n(\theta, \psi)|, \end{aligned}$$

which shows the second result. □

Lemma B.14. *Recall $Z_n(\theta, \psi) = d(\psi, \Psi_n(\theta)) - d(\psi, T_n(\theta))$ for $(\theta, \psi) \in \partial\Theta_I \times \overline{B_M(0)}$ as defined in Lemma B.13. For $Z_n(\theta, \psi)$ defined on the compact local graph $\partial\Theta_I \times \overline{B_M(0)}$ and define its limit counterpart on the limit graph $\partial\Theta_I \times \mathbb{R}^{d_\theta}$ as*

$$Z(\theta, \psi) := d(\psi, \Psi(\theta)) - d(\psi, T(\theta)), \tag{B.29}$$

where $\Psi(\theta)$ and $T(\theta)$ are as in Lemma B.12. Finally for any $\theta \in \partial\Theta_I$ define

$$V(\theta) := d_H(\Psi(\theta), T(\theta)). \tag{B.30}$$

Then it holds that

$$\sup_{(\theta,\psi) \in \partial\Theta_I \times \mathbb{R}^{d_\theta}} |Z(\theta, \psi)| = \sup_{\theta \in \partial\Theta_I} V(\theta).$$

Proof. Note that $\sup_{\psi \in \mathbb{R}^{d_\theta}} |Z(\theta, \psi)| = d_H(\Psi(\theta), T(\theta)) = V(\theta)$ for any $\theta \in \partial\Theta_I$ by the property of d_H; taking the supremum over $\theta \in \partial\Theta_I$ yields the result. □

Lemma B.15. *Let $\theta \in \partial\Theta_I$ and $v \in \mathbb{R}^{d_\theta}$ such that $G_{\mathcal{A}(\theta)}(\theta)v = -D_{\mathcal{A}(\theta)}(\theta)$ (v may not be unique as $v = v' + ker(G_{\mathcal{A}(\theta)}(\theta))$ with v' such solution). Then*

$$\begin{aligned} V(\theta) = d_H\left(\Psi(\theta), T(\theta)\right) &= \sup_{u \in N(\theta), \|u\|=1} |\langle u, v \rangle| \\ &= \sup\left\{ \left| \alpha^t D_{\mathcal{A}(\theta)}(\theta) \right| : \alpha \geq 0, \left\| \alpha^t G_{\mathcal{A}(\theta)}(\theta) \right\| = 1 \right\}. \end{aligned} \tag{B.31}$$

Moreover,

$$V(\theta) = \sup_{\psi \in \mathbb{R}^{d_\theta}} |d(\psi, \Psi(\theta)) - d(\psi, T(\theta))| = \max_{\psi \in \mathbb{R}^{d_\theta} : \psi \in \{0, v\}} |d(\psi, \Psi(\theta)) - d(\psi, T(\theta))|. \tag{B.32}$$

Proof. It follows from Lemma B.12 that $\Psi(\theta)$ and $T(\theta)$ are intersections of half spaces, implying that these sets are convex polyhedral cones (note that these sets $\Psi(\theta)$ and $T(\theta)$ capture local nature of Θ_I at $\theta \in \partial\Theta_I$ and this does not require convexity of Θ_I).

First note that $\Psi(\theta) = T(\theta) + v$ from Lemma B.12. Hence $\sup_{a\in\Psi(\theta)} d(a, T(\theta)) = \sup_{t\in T(\theta)} d(t + v, T(\theta))$. Since $d(t + v, T(\theta)) \leq d(t, T(\theta)) + d(v, T(\theta)) = d(v, T(\theta))$ for any $t \in T(\theta)$, it holds that $\sup_{a\in\Psi(\theta)} d(a, T(\theta)) = d(v, T(\theta))$. In a similar manner, it holds that $\sup_{b\in T(\theta)} d(b, \Psi(\theta)) = \sup_{b\in T(\theta)} d(b, T(\theta) + v) = \sup_{b\in T(\theta)} d(b - v, T(\theta))$; since $d(b - v, T(\theta)) \leq d(b, T(\theta)) + d(-v, T(\theta)) = d(-v, T(\theta))$ for any $b \in T(\theta)$, one obtains $\sup_{b\in T(\theta)} d(b, \Psi(\theta)) = d(-v, T(\theta))$. Combining both quantities leads to

$$\begin{aligned} d_H\left(\Psi(\theta), T(\theta)\right) &= \max\left\{\sup_{a\in\Psi(\theta)} d(a, T(\theta)),\ \sup_{b\in T(\theta)} d(b, \Psi(\theta))\right\} \\ &= \max\left\{d(v, T(\theta)),\ d(-v, T(\theta))\right\}. \end{aligned}$$

Now since $d_H(\Psi(\theta), T(\theta)) = \sup_{\psi\in\mathbb{R}^{d_\theta}} |d(\psi, \Psi(\theta) - d(\psi, T(\theta))|$, putting $\psi = v \in \Psi(\theta)$ or $0 \in T(\theta)$ leads to $d(v, T(\theta))$ or $d(0, \Psi(\theta)) = d(-v, T(\theta))$; that is, $\psi \in \{0, v\}$ achieves $d_H(\Psi(\theta), T(\theta))$, which shows (B.32).

Next I use the result (the "minimum norm duality theorem" (e.g. Luenberger (1969) section 5.13)) $d(x, T) = \sup\{\langle x, u\rangle - s(u, T) : \|u\| \leq 1\}$, where $x \in \mathbb{R}^{d_\theta}$, $T \subset \mathbb{R}^{d_\theta}$ is a closed convex set and $s(u, T) := \sup_{t\in T}\langle u, t\rangle$ (i.e. support function of T). In particular, for closed convex cone $T(\theta)$, it follows that $s(u, T(\theta)) = 0$ if $u \in T(\theta)^o := \{\psi \in \mathbb{R}^{d_\theta} : \langle\psi, t\rangle \leq 0,\ \text{for any } t \in T(\theta)\}$ (polar cone of $T(\theta)$), and $= +\infty$ otherwise. Here note that $T(\theta) = \{\psi \in \mathbb{R}^{d_\theta} : \langle\phi_j(\theta), \psi\rangle \leq 0,\ j \in \mathcal{A}(\theta)\}$ implies $T(\theta)^o = \{\sum_{j\in\mathcal{A}(\theta)} \alpha_j \cdot \phi_j(\theta) : \alpha_j \geq 0\} =: N(\theta)$. Thus,

$$d(x, T(\theta)) = \sup\left\{\langle x, u\rangle : u \in N(\theta), \|u\| \leq 1\right\}.$$

Combining this with the representation of $d_H(\Psi(\theta), T(\theta))$ shows

$$\begin{aligned} d_H\left(\Psi(\theta), T(\theta)\right) &= \max\left\{d(v, T(\theta)),\ d(-v, T(\theta))\right\} \\ &= \max\left\{\sup_{u\in N(\theta):\|u\|\leq 1} \langle v, u\rangle,\ \sup_{u\in N(\theta):\|u\|\leq 1} \langle -v, u\rangle\right\} \\ &= \sup_{u\in N(\theta):\|u\|\leq 1} |\langle v, u\rangle| \\ &= \sup_{u\in N(\theta):\|u\|=1} |\langle v, u\rangle|. \end{aligned}$$

Finally putting $u^t = \alpha^t G_{\mathcal{A}(\theta)}(\theta)$ with $\|\alpha^t G_{\mathcal{A}(\theta)}(\theta)\| = 1$ and noting that $G_{\mathcal{A}(\theta)}(\theta)v = -D_{\mathcal{A}(\theta)}(\theta)$ give

$$\langle v, u\rangle = v^t G_{\mathcal{A}(\theta)}(\theta)^t\alpha = -D_{\mathcal{A}(\theta)}(\theta)^t\alpha,$$

which shows the result in (B.31). □

B.6.2. *Proofs for Proposition B.9.*

Lemma B.16. *Given any subset $\mathcal{I} \subset \mathcal{J}_2$, let $F_{\mathcal{I}} = \{\theta \in \partial\Theta_I : \mathcal{A}(\theta) = \mathcal{I}\}$. Then*

$$\partial\Theta_I = \bigsqcup_{\mathcal{I}\subset\mathcal{J}_2} F_{\mathcal{I}},$$

where the number of disjoint unions is finite.

Proof. This comes from the relation that $|\mathcal{J}_2| \le d_q < \infty$ (indeed $|\mathcal{J}_2| = d_q$ as $\mathcal{J}_1 = \emptyset$ is assumed here). □

Lemma B.17. *For every $\alpha \in \mathbb{R}^{|\mathcal{A}(\theta)|}$ satisfying*

$$\alpha \ge 0, \quad \left\|\alpha^t G_{\mathcal{A}}(\theta)\right\| = 1,$$

it holds that $\|\alpha\| \le 1/c$, where c is the lower bound in Assumption 5.4.

Proof. Pick $\alpha \ge 0$ with $\|\alpha^t G_{\mathcal{A}}(\theta)\| = 1$, and let $s = \sum_{j\in[|\mathcal{A}(\theta)|]} \alpha_j$; by Assumption 5.4 it holds that $s > 0$. Then $\alpha^t G_{\mathcal{A}}(\theta) = s \cdot w$ where $w = \sum_{j\in[|\mathcal{A}(\theta)|]} (\alpha_j/s)\phi_j$, which is in $\mathrm{conv}\{\phi_j(\theta) : j \in \mathcal{A}(\theta)\}$. Thus $\|w\| \ge c$ by Assumption 5.4 and $1 = \|\alpha^t G_{\mathcal{A}}(\theta)\| = s\|w\| \ge s \cdot c$. Since $\|\alpha\|^2 \le (\sum_{j\in[|\mathcal{A}(\theta)|]} \alpha_j)^2$ noting that $\alpha \ge 0$, one obtains $\|\alpha\| \le \sum_{j\in[|\mathcal{A}(\theta)|]} \alpha_j = s \le 1/c$. □

Lemma B.18. *Given an index set $\mathcal{I}$, a $|\mathcal{I}| \times d_\theta$-matrix $G \in \mathbb{R}^{|\mathcal{I}|\times d_\theta}$ and $D \in \mathbb{R}^{|\mathcal{I}|}$, let*

$$F(G, D) := \sup\left\{\left|\alpha^t D\right| : \alpha \ge 0,\ \left\|\alpha^t G\right\| = 1\right\}. \tag{B.33}$$

In particular, this represents $V(\theta) = F(G_{\mathcal{A}(\theta)}(\theta), D_{\mathcal{A}(\theta)}(\theta))$ which is defined in Lemma B.15. For any $G \in \mathbb{R}^{|\mathcal{I}|\times d_\theta}$ such that $d(0, \mathrm{conv}\{\text{rows of } G\}) \ge c > 0$ for some $c > 0$ and for any $D_1, D_2 \in \mathbb{R}^{|\mathcal{I}|}$, we have

$$|F(G, D_1) - F(G, D_2)| \le \frac{1}{c}\|D_1 - D_2\|.$$

Proof. Note that

$$\begin{aligned} |F(G, D_1) - F(G, D_2)| &\le \sup\left\{\left|\alpha^t(D_1 - D_2)\right| : \alpha \ge 0,\ \left\|\alpha^t G\right\| = 1\right\} \\ &\le \|\alpha\| \cdot \|D_1 - D_2\| \\ &\le \frac{1}{c}\|D_1 - D_2\|, \end{aligned}$$

where the last inequality follows by Lemma B.17. □

Lemma B.19. *For a continuous function $D : \partial\Theta_I \mapsto \mathbb{R}^{d_q}$, let*

$$\Gamma(D) := \max_{\mathcal{I}} \sup_{\theta\in\overline{F_{\mathcal{I}}}} F(G_{\mathcal{I}}(\theta), D_{\mathcal{I}}(\theta)), \tag{B.34}$$

where $F_{\mathcal{I}}$ is as in Lemma B.16 and F is as in Lemma B.18. Then $\sup_{\theta\in\partial\Theta_I} V(\theta) = \Gamma(D)$ P-a.s., and

$$|\Gamma(D) - \Gamma(D')| \le \frac{1}{c} \cdot \sup_{\theta\in\partial\Theta_I} \|D(\theta) - D'(\theta)\|.$$

Proof. Recall that $V(\theta) = F(G_{\mathcal{A}(\theta)}(\theta), D_{\mathcal{A}(\theta)}(\theta))$ by Lemma B.15. Since any $\theta \in \partial\Theta_I$ lies in $F_{\mathcal{A}(\theta)} \subset \overline{F_{\mathcal{A}(\theta)}}$ by Lemma B.16, it holds that $\sup_{\theta\in\partial\Theta_I} V(\theta) \leq \Gamma(D)$. For the reverse inequality, let $\theta \in \overline{F_{\mathcal{I}}}$. By continuity of $\Phi_j(\cdot, P)$, $\Phi_j(\theta, P) = 0$ for $j \in \mathcal{I}$, so $\mathcal{I} \subset \mathcal{A}(\theta)$. Any $\alpha \geq 0$ with $\|\alpha^t G_{\mathcal{I}}(\theta)\| = 1$, extended by zeros to $\tilde{\alpha} \in \mathbb{R}^{|\mathcal{A}(\theta)|}$, satisfies $\tilde{\alpha} \geq 0$, $\|\tilde{\alpha}^t G_{\mathcal{A}(\theta)}(\theta)\| = 1$ and $\tilde{\alpha}^t D_{\mathcal{A}(\theta)}(\theta) = \alpha^t D_{\mathcal{I}}(\theta)$. Hence $F(G_{\mathcal{I}}(\theta), D_{\mathcal{I}}(\theta)) \leq V(\theta)$, and $\Gamma(D) \leq \sup_{\theta\in\partial\Theta_I} V(\theta)$. Here D in (27) is continuous P-a.s. by Assumptions 4.9 (ii), (iv) and 5.1 (ii).

For the second result, fix $\mathcal{I}$ and $\theta \in \overline{F_{\mathcal{I}}}$. Then

$$|F(G_{\mathcal{I}}(\theta), D_{\mathcal{I}}(\theta)) - F(G_{\mathcal{I}}(\theta), D'_{\mathcal{I}}(\theta)| \leq \frac{1}{c} \|D(\theta) - D'(\theta)\|,$$

by Lemma B.18. Since

$$\begin{aligned}
&\left|\max_{\mathcal{I}} \sup_{\theta\in\overline{F}_{\mathcal{I}}} F(G_{\mathcal{I}}(\theta), D_{\mathcal{I}}(\theta)) - \max_{\mathcal{I}} \sup_{\theta\in\overline{F}_{\mathcal{I}}} F(G_{\mathcal{I}}(\theta), D'_{\mathcal{I}}(\theta))\right| \\
&\qquad \leq \max_{\mathcal{I}} \left|\sup_{\theta\in\overline{F}_{\mathcal{I}}} F(G_{\mathcal{I}}(\theta), D_{\mathcal{I}}(\theta)) - \sup_{\theta\in\overline{F}_{\mathcal{I}}} F(G_{\mathcal{I}}(\theta), D'_{\mathcal{I}}(\theta))\right| \\
&\qquad \leq \max_{\mathcal{I}} \sup_{\theta\in\overline{F}_{\mathcal{I}}} |F(G_{\mathcal{I}}(\theta), D_{\mathcal{I}}(\theta)) - F(G_{\mathcal{I}}(\theta), D'_{\mathcal{I}}(\theta)|,
\end{aligned}$$

the second result follows. □

Lemma B.20. *Let $\theta \in \partial\Theta_I$ and $\mathcal{I} \subset \mathcal{A}(\theta)$ non-empty. For $j \in \mathcal{I}$ let $s_j(\theta) = -\sqrt{n} \cdot \Phi_j(\theta, P) \geq 0$ and $s = (s_j(\theta))_{j\in\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$. For vector $D_{\mathcal{I}} = (D_j)_{j\in\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$, define*

$$\begin{aligned}
T^s(\theta) &:= \left\{\psi \in \mathbb{R}^{d_\theta} : \langle \phi_j(\theta), \psi\rangle \leq s_j(\theta),\ j \in \mathcal{I}\right\}, \\
\Psi^s(\theta) &:= \left\{\psi \in \mathbb{R}^{d_\theta} : \langle \phi_j(\theta), \psi\rangle \leq -D_j(\theta) + s_j(\theta),\ j \in \mathcal{I}\right\}.
\end{aligned}$$

Then $G_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|\times d_\theta}$ is of full row rank, implying that there exist $v, w \in \mathbb{R}^{d_\theta}$ such that $G_{\mathcal{I}}(\theta)v = -D_{\mathcal{I}}(\theta)$ and $G_{\mathcal{I}}(\theta)w = s$. Moreover, it follows that $\Psi^s(\theta) = T^s(\theta) \oplus v$ and $T^s(\theta) = T^0(\theta) \oplus w$. Hence,

$$\sup_{\psi\in\overline{B_M(0)}} |d(\psi, \Psi^s(\theta)) - d(\psi, T^s(\theta))| = d_H(\Psi^s(\theta), T^s(\theta)) = d_H\left(T^0(\theta) \oplus v, T^0(\theta)\right).$$

Proof. Assumption 5.4 (iii) implies that $G_{\mathcal{A}(\theta)}(\theta)$ are of full row rank; since $G_{\mathcal{I}}(\theta)$ is a submatrix of $G_{\mathcal{A}(\theta)}(\theta)$, $G_{\mathcal{I}}(\theta)$ has also full row rank. The first results are then immediate. For the second results, note that

$$\begin{aligned}
\Psi^s(\theta) &= \left\{\psi \in \mathbb{R}^{d_\theta} : G_{\mathcal{I}}(\theta)\psi \leq -D_{\mathcal{I}}(\theta) + s\right\} \\
&= \left\{\psi \in \mathbb{R}^{d_\theta} : G_{\mathcal{I}}(\theta)(\psi - v) \leq s\right\} \\
&= T^s(\theta) \oplus v,
\end{aligned}$$

and likewise,

$$T^s(\theta) = \left\{\psi \in \mathbb{R}^{d_\theta} : G_{\mathcal{I}}(\theta)\psi \leq s\right\} = \left\{\psi \in \mathbb{R}^{d_\theta} : G_{\mathcal{I}}(\theta)(\psi - w) \leq 0\right\} = T^0(\theta) \oplus w.$$

For the last equivalence, since the Hausdorff distance is invariant for shift transformation,

$$d_H(\Psi^s(\theta), T^s(\theta)) = d_H(T^0(\theta) \oplus w \oplus v, T^0(\theta) \oplus w) = d_H(T^0(\theta) \oplus v, T^0(\theta)).$$

□

Lemma B.21. *Let* $D_n(\theta) := \sqrt{n} \cdot (\Phi(\theta, \mathbb{P}_n) - \Phi(\theta, P))$. *Then*

$$\sqrt{n} \cdot d_H(\widehat{\Theta}_I, \Theta_I) = \Gamma(D_n) + o_P(1).$$

Proof. First let

$$V_n(\theta) := d_H(\Psi_n(\theta), T_n(\theta)), \quad \theta \in \partial\Theta_I, \tag{B.35}$$

which can be seen as the sample counterpart of $V(\theta)$ in (B.30). As in Lemma B.14, one can get

$$\sup_{(\theta,\psi)\in\partial\Theta_I \times \overline{B_M(0)}} |Z_n(\theta, \psi)| = \sup_{\theta\in\partial\Theta_I} V_n(\theta).$$

Since it follows with probability approaching 1 that

$$\sqrt{n} \cdot \left(d_H(\widehat{\Theta}_I, \Theta_I)\right) = \sup_{x\in\mathcal{T}_n(M)} |Z_n(x)| = \sup_{(\theta,\psi)\in\partial\Theta_I \times \overline{B_M(0)}} |Z_n(\theta, \psi)|,$$

where the first equality follows from Lemma B.11 and the second one follows from Lemma B.13, it holds that $\sqrt{n} \cdot (d_H(\widehat{\Theta}_I, \Theta_I)) = \sup_{\theta\in\partial\Theta_I} V_n(\theta)$.

Next note that $D_n(\theta) \rightsquigarrow D(\theta) = (D_j(\theta))_{j\in\mathcal{J}_2}$ in $\ell^\infty(\Theta)$ by the Donsker property in Assumption 5.1 and $\Phi(\cdot, P) \in C^2(\Theta)$ in Assumption 5.4. By the Skorokhod–Dudley–Wichura construction (van der Vaart and Wellner (2023, Theorem 1.10.3)), we may assume that $\sup_{\theta\in\partial\Theta_I} \|D_n(\theta) - D(\theta)\| \to 0$ almost surely, and I argue on a fixed path on which this holds. On it, $\Gamma(D_n) = \Gamma(D) + o(1)$ by Lemma B.19. Here recall that $\Gamma(D) = \sup_{\theta\in\partial\Theta_I} V(\theta)$ as seen in Lemma B.19. Combining the arguments, in order to prove the result in Lemma B.21 it suffices to show

$$\limsup_{n\to\infty} \sup_{\theta\in\partial\Theta_I} V_n(\theta) \le \sup_{\theta\in\partial\Theta_I} V(\theta) \le \liminf_{n\to\infty} \sup_{\theta\in\partial\Theta_I} V_n(\theta).$$

For the second inequality, note that

$$|Z_n(\theta, \psi)| \le \sup_{\theta\in\partial\Theta_I} V_n(\theta)$$

for any fixed $(\theta, \psi) \in \partial\Theta_I \times \overline{B_M(0)}$. Note also that

$$\begin{aligned} Z_n(\theta, \psi) &= d(\psi, \Psi_n(\theta)) - d(\psi, T_n(\theta)) \\ &= d(\psi, \Psi(\theta)) - d(\psi, T(\theta)) + o(1) = Z(\theta, \psi) + o(1), \end{aligned}$$

where the first equality comes from the proof in Lemma B.13, the second one comes from Lemma B.12 combined with the fact that the metric $d(\psi, \cdot)$ is continuous, and the third one follows by the definition of $Z(\theta, \psi)$ in (B.29). Hence one gets $|Z(\theta, \psi)| \le \liminf_{n\to\infty} \sup_{\theta\in\partial\Theta_I} V_n(\theta)$, and the second inequality follows by taking the supremum over (θ, ψ) and using Lemma B.14.

For the first inequality, one can prove it by contradiction. Suppose that the first inequality would not hold. Then there exist some $\epsilon > 0$ and sequences $\{\theta_n\}_{n\in\mathbb{N}} \subset \partial\Theta_I$ and $\{\psi_n\}_{n\in\mathbb{N}} \subset \overline{B_M(0)}$ such that $|Z_n(\theta_n, \psi_n)| > \sup_{\theta\in\partial\Theta_I} V(\theta) + \epsilon$. Recall $s_j(\theta) = -\sqrt{n}\cdot\Phi_j(\theta, P)$ defined in Lemma B.20; these take values in $[0,\infty]$ which is compact with the standard ordered topology. Since $\partial\Theta_I$, $\overline{B_M(0)}$ and $[0,\infty]$ are compact, one can take a subsequence such that $\theta_{n_k} \to \theta^* \in \partial\Theta_I$, $\psi_{n_k} \to \psi^* \in \overline{B_M(0)}$ and $s_j(\theta_{n_k}) \to s_j^* \in [0,\infty]$. Now let $\mathcal{I} := \{j \in \mathcal{J}_2 : s_j^* < \infty\}$; clearly $\mathcal{I} \subset \mathcal{A}(\theta^*)$. Define

$$\begin{aligned}
T^{s^*} &:= \{\psi \in \mathbb{R}^{d_\theta} : G_{\mathcal{I}}(\theta^*)\psi \le s_{\mathcal{I}}^*\},\\
&= \{\psi \in \mathbb{R}^{d_\theta} : \langle \phi_j(\theta^*), \psi\rangle \le s_j^*,\ j \in \mathcal{I}\},\\
\Psi^{s^*} &:= \{\psi \in \mathbb{R}^{d_\theta} : G_{\mathcal{I}}(\theta^*)\psi \le -D_{\mathcal{I}}(\theta^*) + s_{\mathcal{I}}^*\},\\
&= \{\psi \in \mathbb{R}^{d_\theta} : \langle \phi_j(\theta^*), \psi\rangle \le -D_j(\theta^*) + s_j^*,\ j \in \mathcal{I}\},
\end{aligned}$$

which are the slack-local sets as in Lemma B.20. Similar to Lemma B.12, the Taylor expansion of $\Phi_j(\cdot, \mathbb{P}_{n_k})$ around $\{\theta_{n_k} + \psi/\sqrt{n_k}\}$ shows (by the mean value theorem, since $\Phi(\cdot, P)$ is C^1 by Assumption 5.4 (i), and by the continuity of D at θ^*)

$$\sqrt{n_k}\cdot\Phi_j(\theta_{n_k} + \psi/\sqrt{n_k}, \mathbb{P}_{n_k}) = -s_j(\theta_{n_k}) + \langle \phi_j(\theta^*), \psi\rangle + D_j(\theta^*) + o(1),$$

which holds uniformly over $\psi \in \overline{B_M(0)}$. Applying the similar logic as in Lemma B.12 implies that $\Psi_n(\theta_{n_k}) \cap \overline{B_M(0)} \rightsquigarrow \Psi^{s^*} \cap \overline{B_M(0)}$ and $T_n(\theta_{n_k}) \cap \overline{B_M(0)} \to T^{s^*} \cap \overline{B_M(0)}$. Accordingly, it follows that

$$\begin{aligned}
|Z_{n_k}(\theta_{n_k}, \psi_{n_k})| &= \left|d(\psi^*, \Psi^{s^*}) - d(\psi^*, T^{s^*})\right| + o(1)\\
&\le \sup_{\psi^*\in\overline{B_M(0)}} \left|d(\psi^*, \Psi^{s^*}) - d(\psi^*, T^{s^*})\right| + o(1)\\
&\le d_H(T^{0^*} + v, T^{0^*}) + o(1)\\
&= F(G_{\mathcal{I}}(\theta^*), D_{\mathcal{I}}(\theta^*)) + o(1),
\end{aligned}$$

where $T^{0^*} = \{\psi \in \mathbb{R}^{d_\theta} : G_{\mathcal{I}}(\theta^*)\psi \le 0\}$ and $\Psi^{0^*} = \{\psi \in \mathbb{R}^{d_\theta} : G_{\mathcal{I}}(\theta^*)\psi \le -D_{\mathcal{I}}(\theta^*)\} = T^{0^*} \oplus v$ with $G_{\mathcal{I}}(\theta^*)v = -D_{\mathcal{I}}(\theta^*)$. Since $F(G_{\mathcal{I}}(\theta^*), D_{\mathcal{I}}(\theta^*)) = V(\theta^*)$ as seen in Lemma B.18, this further implies $V(\theta^*) > \sup_{\theta\in\partial\Theta_I} V(\theta) + \epsilon$ which is a contradiction. □

Proof. Proposition B.9.
Lemma B.21 shows that $\sqrt{n}\cdot d_H(\widehat{\Theta}_I, \Theta_I) = \Gamma(D_n) + o_P(1)$. Since $\Gamma : \ell^\infty(\partial\Theta_I) \mapsto \mathbb{R}_+$ is continuous as seen in Lemma B.19 and that $D_n \rightsquigarrow D$ in $\ell^\infty(\partial\Theta_I)$ by Assumption 5.1 (ii) and Assumption 5.4 (i), one then gets $\Gamma(D_n) \rightsquigarrow \Gamma(D)$ by the continuous mapping theorem. It then follows that

$$\sqrt{n}\cdot d_H(\widehat{\Theta}_I, \Theta_I) \rightsquigarrow \Gamma(D) = \sup_{\theta\in\partial\Theta_I} V(\theta) = \sup_{(\theta,\psi)\in\partial\Theta_I\times\mathbb{R}^{d_\theta}} |Z(\theta, \psi)|,$$

where the first equality follows from Lemma B.19 and the second one from Lemma B.14. This shows the desired result. □

References

Aghion, Philippe et al. (2005). "Competition and Innovation: An Inverted-U Relationship". *The Quarterly Journal of Economics* 120(2), pp. 701–728.

Allcott, Hunt and Nathan Wozny (2014). "Gasoline Prices, Fuel Economy, and the Energy Paradox". *The Review of Economics and Statistics* 96(5), pp. 779–795.

Andrews, Donald W. K. (1999). "Estimation When a Parameter Is on a Boundary". *Econometrica* 67(6), pp. 1341–1383.

Andrews, Donald W. K. and Patrik Guggenberger (2009). "Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities". *Econometric Theory* 25(3), pp. 669–709.

Andrews, Donald W. K. and Gustavo Soares (2010). "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection". *Econometrica* 78(1), pp. 119–157.

Andrews, Isaiah and Anna Mikusheva (2022). "Optimal Decision Rules for Weak GMM". *Econometrica* 90(2), pp. 715–748.

Armstrong, Timothy B. (2014). "Weighted KS Statistics for Inference on Conditional Moment Inequalities". *Journal of Econometrics* 181(2), pp. 92–116.

Ashenfelter, Orley and Michael Greenstone (2004). "Using Mandated Speed Limits to Measure the Value of a Statistical Life". *Journal of Political Economy* 112(S1), S226–S267.

Auten, Gerald and David Splinter (2024). "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends". *Journal of Political Economy* 132(7), pp. 2179–2227.

Baumeister, Christiane and James D. Hamilton (2015). "Sign Restrictions, Structural Vector Autoregressions, and Useful Prior Information". *Econometrica* 83(5), pp. 1963–1999.

Beresteanu, Arie, Ilya Molchanov, and Francesca Molinari (2011). "Sharp Identification Regions in Models with Convex Moment Predictions". *Econometrica* 79(6), pp. 1785–1821.

Beresteanu, Arie and Francesca Molinari (2008). "Asymptotic Properties for a Class of Partially Identified Models". *Econometrica* 76(4), pp. 763–814.

Bickel, Peter J. et al. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer New York.

Bond, Steve et al. (2021). "Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation from Production Data". *Journal of Monetary Economics* 121, pp. 1–14.

Bontemps, Christian, Thierry Magnac, and Eric Maurin (2012). "Set Identified Linear Models". *Econometrica* 80(3), pp. 1129–1155.

Borwein, Jonathan and Adrian Lewis (2006). *Convex Analysis and Nonlinear Optimization*. Springer New York.

Bugni, Federico A. (2010). "Bootstrap Inference in Partially Identified Models Defined by Moment Inequalities: Coverage of the Identified Set". *Econometrica* 78(2), pp. 735–753.

Bugni, Federico A., Ivan A. Canay, and Patrik Guggenberger (2012). "Distortions of Asymptotic Confidence Size in Locally Misspecified Moemnt Inequality Models". *Econometrica* 80(4), pp. 1741–1768.

Busse, Meghan R., Christopher R. Knittel, and Florian Zettelmeyer (2013). "Are Consumers Myopic? Evidence from New and Used Car Purchases". *American Economic Review* 103(1), pp. 220–256.

Canay, Ivan A. (2010). "EL Inference for Partially Identified Models: Large Deviations Optimality and Bootstrap Validity". *Journal of Econometrics* 156(2), pp. 408–425.

Canova, Fabio and Luca Sala (2009). "Back to Square One: Identification Issues in DSGE Models". *Journal of Monetary Economics* 56(4), pp. 431–449.

Chamberlain, Gary (1987). "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions". *Journal of Econometrics* 34(3), pp. 305–334.

Chen, Xiaohong, Timothy M. Christensen, and Elie Tamer (2018). "Monte Carlo Confidence Sets for Identified Sets". *Econometrica* 86(6), pp. 1965–2018.

Chernozhukov, Victor, Han Hong, and Elie Tamer (2007). "Estimation and Confidence Regions for Parameter Sets in Econometric Models". *Econometrica* 75(5), pp. 1243–1284.

Chetty, Raj (2008). "Moral Hazard versus Liquidity and Optimal Unemployment Insurance". *Journal of Political Economy* 116(2), pp. 173–234.

– (2009). "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods". *Annual Review of Economics* 1, pp. 451–488.

Chetty, Raj, Adam Looney, and Kory Kroft (2009). "Salience and Taxation: Theory and Evidence". *American Economic Review* 99(4), pp. 1145–1177.

De Bartolome, Charles A. M. (1995). "Which Tax Rate Do People Use: Average or Marginal?" *Journal of Public Economics* 56(1), pp. 79–96.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger (2020). "The Rise of Market Power and the Macroeconomic Implications". *The Quarterly Journal of Economics* 135(2), pp. 561–644.

De Loecker, Jan and Frederic Warzynski (2012). "Markups and Firm-Level Export Status". *American Economic Review* 102(6), pp. 2437–2471.

Diamond, Peter and Emmanuel Saez (2011). "The Case for a Progressive Tax: From Basic Research to Policy Recommendations". *Journal of Economic Perspectives* 25(4), pp. 165–190.

Fang, Zheng and Andres Santos (2019). "Inference on Directionally Differentiable Functions". *The Review of Economic Studies* 86(1), pp. 377–412.

Farhi, Emmanuel and Xavier Gabaix (2020). "Optimal Taxation with Behavioral Agents". *American Economic Review* 110(1), pp. 298–336.

Federer, Herbert (1959). "Curvature Measures". *Transactions of the American Mathematical Society* 93(3), pp. 418–491.

Feldman, Naomi E., Peter Katuščák, and Laura Kawano (2016). "Taxpayer Confusion: Evidence from the Child Tax Credit". *American Economic Review* 106(3), pp. 807–835.

Giacomini, Raffaella and Toru Kitagawa (2021). "Robust Bayesian Inference for Set-Identified Models". *Econometrica* 89(4), pp. 1519–1556.

Giné, Evarist, Marjorie G. Hahn, and Joel Zinn (1983). "Limit Theorems for Random Sets: An Application of Probability in Banach Space Results". *Probability in Banach Spaces IV (Proceedings of the Seminar Held in Oberwolfach, FRG, July 1982)*. Springer Berlin, Heidelberg.

Granziera, Elenora, Hyungsik Roger Moon, and Frank Schorfheide (2018). "Inference for VARs Identified with Sign Restrictions". *Quantitative Economics* 9(3), pp. 1087–1121.

Gruber, Jon and Emmanuel Saez (2002). "The Elasticity of Taxable Income: Evidence and Implications". *Journal of Public Economics* 84(1), pp. 1–32.

Gruber, Jonathan (1997). "The Consumption Smoothing Benefits of Unemployment Insurance". *American Economic Review* 87(1), pp. 192–205.

Hájek, Jaroslav (1970). "A Characterization of Limiting Distributions of Regular Estimates". *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 14, pp. 323–330.

– (1972). "Local Asymptotic Minimax and Admissibility in Estimation". *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* 1, pp. 175–194.

Hansen, Lars Peter (1982). "Large Sample Properties of Generalized Method of Moments Estimators". *Econometrica* 50(4), pp. 1029–1054.

Hendren, Nathaniel and Ben Sprung-Keyser (2020). "A Unified Welfare Analysis of Government Policies". *The Quarterly Journal of Economics* 135(3), pp. 1209–1318.

Hirano, Keisuke and Jack R. Porter (2009). "Asymptotics for Statistical Treatment Rules". *Econometrica* 77(5), pp. 1683–1701.

– (2012). "Impossibility Results for Nondifferentiable Functionals". *Econometrica* 80(4), pp. 1769–1790.

– (2020). "Chapter 4 - Asymptotic Analysis of Statistical Decision Rules in Econometrics". *Handbook of Econometrics, Volume 7A*. Vol. 7. Elsevier, pp. 283–354.

Kaido, Hiroaki (2016). "A Dual Approach to Inference for Partially Identified Econometric Models". *Journal of Econometrics* 192(1), pp. 269–290.

Kaido, Hiroaki, Francesca Molinari, and Jörg Stoye (2022). "Constraint Qualifications in Partial Identification". *Econometric Theory* 38(3), pp. 596–619.

Kaido, Hiroaki and Andres Santos (2014). "Asymptotically Efficient Estimation of Models Defined by Convex Moment Inequalities". *Econometrica* 82(1), pp. 387–413.

Komunjer, Ivana and Serena Ng (2011). "Dynamic Identification of Dynamic Stochastic General Equilibrium Models". *Econometrica* 79(6), pp. 1995–2032.

Lardeux, Raphaël (2023). "Behavioral Cross-Influence of a Shadow Tax Bracket: Evidence from Bunching where Income Tax Liabilities Start". *Journal of Public Economics* 226, p. 104950.

Le Cam, Lucien (1972). "Limits of Experiments". *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* 1, pp. 245–261.

Lee, John M. (2012). *Introduction to Smooth Manifolds*. Springer New York.

Luenberger, David G. (1969). *Optimization by Vector Space Methods*. John Wiley & Sons.

Manski, Charles F. (1989). "Anatomy of the Selection Problem". *Journal of Human Resources* 24(3), pp. 343–360.

Manski, Charles F. and Elie Tamer (2002). "Inference on Regressions with Interval Data on a Regressor or Outcome". *Econometrica* 70(2), pp. 519–546.

Molchanov, Ilya (2017). *Theory of Random Sets*. Springer London.

Molinari, Francesca (2020). "Microeconometrics with Partial Identification". *Handbook of Econometrics*. Ed. by Steven N. Durlauf et al. Vol. 7A. Elsevier, pp. 355–486.

Moon, Hyungsik Roger and Frank Schorfheide (2012). "Bayesian and Frequentist Inference in Partially Identified Models". *Econometrica* 80(2), pp. 755–782.

Müller, Ulrich K. (2011). "Efficient Tests under a Weak Convergence Assumption". *Econometrica* 79(2), pp. 395–435.

Oster, Emily (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence". *Journal of Business & Economic Statistics* 37(2), pp. 187–204.

Piketty, Thomas and Emmanuel Saez (2003). "Income Inequality in the United States, 1913–1998". *The Quarterly Journal of Economics* 118(1), pp. 1–41.

Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman (2018). "Distributional National Accounts: Methods and Estimates for the United States". *The Quarterly Journal of Economics* 133(2), pp. 553–609.

Poliquin, R. A., R. T. Rockafellar, and L. Thibault (2000). "Local Differentiability of Distance Functions". *Transactions of the American Mathematical Society* 352(11), pp. 5231–5249.

Rees-Jones, Alex and Dmitry Taubinsky (2020). "Measuring "Schmeduling"". *The Review of Economic Studies* 87(5), pp. 2399–2438.

Rockafellar, R. Tyrrell and Roger J. B. Wets (1998). *Variational Analysis*. Springer Berlin.

Romano, Joseph P. and Azeem M. Shaikh (2010). "Inference for the Identified Set in Partially Identified Econometric Models". *Econometrica* 78(1), pp. 169–211.

Saez, Emmanuel (2001). "Using Elasticities to Derive Optimal Income Tax Rates". *The Review of Economic Studies* 68(1), pp. 205–229.

Schneider, Rolf (2013). *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press.

Shapiro, Alexander (1990). "On Concepts of Directional Differentiability". *Journal of Optimization Theory and Applications* 66(3), pp. 477–487.

Taubinsky, Dmitry and Alex Rees-Jones (2018). "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment". *The Review of Economic Studies* 85(4), pp. 2462–2496.

Uhlig, Harald (2005). "What Are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure". *Journal of Monetary Economics* 52(2), pp. 381–419.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Van der Vaart, Aad (1991). "An Asymptotic Representation Theorem". *International Statistical Review* 59(1), pp. 97–121.

Van der Vaart, Aad W. and Jon A. Wellner (2023). *Weak Convergence and Empirical Processes*. Springer Cham.

Wainwright, Martin J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Yildiz, Neşe (2012). "Consistency of Plug-in Estimators of Upper Contour and Level Sets". *Econometric Theory* 28(2), pp. 309–327.