UNIVERSITY OF CALIFORNIA

Los Angeles

Essays in Econometrics

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Economics

by

Kirill Ponomarev

2022

ABSTRACT OF THE DISSERTATION

Essays in Econometrics

by

Kirill Ponomarev
Doctor of Philosophy in Economics
University of California, Los Angeles, 2022
Professor Rosa Liliana Matzkin, Co-Chair
Professor Andres Santos, Co-Chair

The chapters of this dissertation are devoted to three different topics.

The first chapter studies estimation of parameters expressed via non-differentiable functions. Such parameters are abundant in econometric models and typically take the form of maxima or minima of some estimable objects. Examples include bounds on the average treatment effects in non-experimental settings, identified sets for the coefficients in regression models with interval-valued data, bounds on the distribution of wages accounting for selection into employment, and many others. I consider estimators of the form $\phi(\hat{\theta}_n + \hat{v}_{1,n}) + \hat{v}_{2,n}$, where $\hat{\theta}_n$ is the efficient estimator for θ_0, and $\hat{v}_{1,n}, \hat{v}_{2,n}$ are suitable adjustment terms. I characterize the optimal adjustment terms and develop a general procedure to compute them from the data. A simulation study shows that the proposed estimator can have lower finite-sample bias and variance than the existing alternatives. As an application, I consider estimating the bounds on the distribution of valuations and the optimal reserve price in English auctions with independent private values. Empirically calibrated simulations show that the resulting estimates are substantially sharper than the previously available ones.

The second chapter studies inequality selection in partially identified models. Many partially identified models have the following structure: given a parameter vector and covariates, the model produces a set of predictions while the researcher observes a single outcome. Examples include entry games with multiple equilibria, network formation models, discrete-choice models with endogenous explanatory variables or heterogeneous choice sets, and auctions. Sharp identified sets for structural parameters in such models can be characterized via a special kind of moment inequalities. For a given parameter value, the inequalities verify that the observed conditional distribution of the outcome given covariates belongs to the set of distributions admitted by the model. In practice, checking all of the inequalities is often computationally infeasible, and many of them may not even be informative. Therefore, some inequality selection is required. In this chapter, I propose a new analytical criterion that dramatically reduces the number of inequalities required to characterize the sharp identified set. In settings where the outcome space is finite, I characterize the smallest subset of inequalities that guarantees sharpness and show that it can be efficiently computed using graph propagation techniques. I apply the proposed criterion in the context of market entry games, network formation, auctions, and discrete-choice.

The third chapter (coauthored with Liqiang Shi) is about model selection for policy learning. When treatment effects are heterogeneous, a decision maker that has access to (quasi-) experimental data can attempt to find the optimal policy function, mapping observable characteristics into treatment choices, to maximize utilitarian welfare. When several different policy classes are available, the choice of the policy class poses a model selection problem. In this chapter, following Athey and Wager (2021) and Mbakop and Tabord-Meehan (2021), we propose a policy learning algorithm that leverages doubly-robust estimation and incorporates data-driven model selection. We show that the proposed algorithm automatically selects the best available class of policies and achieves the optimal $n^{-1/2}$ rate of convergence in terms of expected regret. We also refine some of the existing related results and derive a new finite-sample lower bound on expected regret.

The dissertation of Kirill Ponomarev is approved.

Denis Nikolaye Chetverikov

Shuyang Sheng

Hyungsik Moon

Andres Santos, Committee Co-Chair

Rosa Liliana Matzkin, Committee Co-Chair

University of California, Los Angeles

2022

To my mother Olga and my father Aleksandr,
for their love, care, and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisors, Rosa Matzkin, Andres Santos, and Denis Chetverikov, for their constant guidance and support. I am grateful for the opportunity to work with them.

I would also like to thank Jinyong Hahn, Zhipeng Liao, Shuyang Sheng, Simon Board, Moritz Meyer-Ter-Vehn, and Tomasz Sadzik, and all participants of the proseminars in Econometrics and Microeconomic Theory, for many enlightening discussions over the years.

Finally, I would like to thank my friends Tomas, Ben, Brian, Diego, Leo, Victoria, and Fernanda, who went through graduate school with me and made this journey truly enjoyable.

VITA

2012–2016	B.S. in Economics, *summa cum laude*, Higher School of Economics, Moscow, Russia
2018	M.A. in Economics, Department of Economics, UCLA.
2018–2021	Teaching Assistant, Department of Economics, UCLA.
2018–2021	Summer Sessions Instructor, Department of Economics, UCLA.

CHAPTER 1

On Efficient Estimation of Directionally Differentiable Functionals

1.1 Introduction

Many econometric models concern parameters of the form $\phi(\theta_0)$, where ϕ is a known function that is directionally but not necessarily fully differentiable, and θ_0 is an unknown but estimable object. Such $\phi(\theta_0)$ may represent, for instance, the bounds on a parameter of interest in a partially-identified model, or a parameter defined as the value function of an optimization problem that may have multiple solutions. Examples include bounds on treatment effects obtained by taking minima or maxima of the estimated conditional moments (e.g., Manski and Pepper, 2000, 2009; Shaikh and Vytlacil, 2011), identified sets for the coefficients in regression models with interval-valued data (Manski and Tamer, 2002; Beresteanu and Molinari, 2008; Bontemps et al., 2012), bounds on the distribution of wages accounting for selection into employment (e.g., Blundell et al., 2007), and bounds on the distribution of valuations and optimal reserve prices derived from the observed distribution of bids in English auctions (Haile and Tamer, 2003; Aradillas-López et al., 2013a; Chesher and Rosen, 2017), among others.[1]

The lack of full differentiability of the function ϕ complicates estimation of such param-

[1]Other examples include bounds on structural parameters in market entry and discrete choice models (Ciliberto and Tamer, 2009; Beresteanu et al., 2011; Pakes et al., 2007, 2015), shape restrictions via projections (Fang, 2018), and the breakdown frontiers in the recent literature on sensitivity analysis (Kline and Santos, 2013; Masten and Poirier, 2020). A more detailed discussion is provided in Section 1.2.2.

eters. Assuming that an efficient estimator $\hat{\theta}_n$ for θ_0 is available, a natural approach is to estimate $\phi(\theta_0)$ with $\phi(\hat{\theta}_n)$. However, the properties of such "plug-in" estimator critically depend on the value of θ_0. If the full differentiability of the function ϕ fails at θ_0, then the "plug-in" estimator will be asymptotically biased (Hirano and Porter, 2012) and inefficient (Song, 2014; Fang, 2018). Moreover, in such cases, one faces a bias-variance trade-off: Since unbiased estimators may not exist, attempting to reduce the bias "too much" may dramatically increase the variance of the resulting estimator (Doss and Sethuraman, 1989). The existing bias-reduction approaches do not take the bias-variance trade-off into account, while the analysis of efficient estimators is very limited.

In this paper, I study efficient estimators for such parameters in a general setting. I assume that the parameter θ_0 is "well-behaved," in the sense that a regular efficient estimator $\hat{\theta}_n$ is available, and that the function ϕ is everywhere directionally differentiable. To accommodate applications such as English auctions or regressions with interval-valued data, I allow both θ_0 and $\phi(\theta_0)$ to take values in finite or infinite-dimensional spaces. I consider a flexible class of estimators of the form

$$\phi\left(\hat{\theta}_n + \frac{\hat{v}_{1,n}}{\sqrt{n}}\right) + \frac{\hat{v}_{2,n}}{\sqrt{n}}, \tag{1.1}$$

where $\hat{v}_{1,n}$, $\hat{v}_{2,n}$ are adjustment terms that converge in probability to constants, and show how to obtain the optimal estimator within this class.

The proposed estimator has two key features. First, it automatically adapts to the presence or failure of full differentiability. That is, if the data suggest that the function ϕ is likely to be fully differentiable at θ_0, both adjustment terms will be equal to zero by construction. In this case, the proposed estimator reduces to $\phi(\hat{\theta}_n)$, which is known to be efficient under full differentiability (e.g. van der Vaart, 1988). On the other hand, if the data reveal that the full differentiability is likely to fail at θ_0, the adjustment terms will differ from zero and improve on the "plug-in" estimator. Second, the optimal adjustment terms depend on the loss function chosen to evaluate and compare different estimators. Under full

differentiability, the "plug-in" estimator $\phi(\hat{\theta}_n)$ is known to be efficient for any symmetric "bowl-shaped" loss function, so that the choice of a particular functional form is irrelevant (e.g. van der Vaart, 1988). In contrast, when differentiability fails, the adjustment terms can depend on the loss function, suggesting that the latter should be tailored to specific applications. In particular, choosing the squared loss function allows to select the adjustment terms to balance the bias-variance trade-off.

In order to accommodate a variety of econometric models and parameters in a tractable way, as a notion of efficiency I employ Local Asymptotic Minimaxity.[2] To elaborate, suppose that the data $X_1, \ldots, X_n$ are an i.i.d. sample with a common distribution $P \in \mathbf{P}$, where $\mathbf{P}$ denotes the model (i.e., the set of all plausible distributions, consistent with the maintained assumptions). Let θ_0 denote some root-n estimable feature of the distribution P, and $\hat{\phi}_n$ denote a generic estimator for the target parameter $\phi(\theta_0)$. Letting l denote a non-negative loss function, the quality of different estimators can be evaluated by their risk, $\mathbb{E}_P\{l(\sqrt{n}(\hat{\phi}_n - \phi(\theta_0)))\}$, where the expectation is calculated with respect to the data distributed according to P.[3] For every fixed n, it is understood that the lower the risk, the better the estimator. The idea of Local Asymptotic Minimaxity is to compare estimators in terms of the asymptotic risk in a locally-worst-case scenario, that is,

$$\liminf_{n\to\infty} \sup_{\tilde{P}\in V_n(P)} \mathbb{E}_{\tilde{P}} \left\{ l\left(\sqrt{n}(\hat{\phi}_n - \phi(\theta(\tilde{P})))\right)\right\}, \tag{1.2}$$

where $V_n(P) \subset \mathbf{P}$ denote certain "local neighborhoods" of P that shrink as n approaches infinity and only contain distributions that are hard to distinguish from P empirically. Any estimator sequence $\{\hat{\phi}_n\}$ that minimizes the above expression is called Locally Asymptoti-

[2]It is worth-noting that, due to the potential lack of full differentiability, regular or unbiased estimators may not exist (van der Vaart, 1991; Hirano and Porter, 2012), and therefore traditional optimality criteria, searching for the "best regular" or "best minimum-variance unbiased" estimators, are inapplicable. Local Asymptotic Minimaxity is applicable more broadly, see Section 1.3.

[3]For example, for a real-valued parameters, the quadratic loss $l(x) = x^2$ corresponds to the mean-squared error, $\mathbb{E}_P\{(\sqrt{n}(\hat{\phi}_n - \phi_0))^2\} = \mathbb{V}ar_P\{\sqrt{n}(\hat{\phi}_n - \phi_0)\} + \{\mathbb{E}_P(\sqrt{n}(\hat{\phi}_n - \phi_0))\}^2$. Note that both the distribution of the estimator $\hat{\phi}_n$ and the value of the target parameter $\phi(\theta_0)$ depend on the distribution P of the data.

cally Minimax (or LAM). A more precise formulation requires substantial background and is discussed in Section 1.3.

To obtain the LAM estimator within the class (1.1), I proceed in two steps. First, I show that the LAM risk, given by (1.2), of any such estimator is bounded from below by

$$\inf_{v_1,v_2} \sup_{s\in S(Z)} \mathbb{E}\left\{ l\left(\phi_0'(Z+v_1+s) - \phi_0'(s) + v_2 \right)\right\}, \tag{1.3}$$

where a random vector (or process) Z denotes the distributional limit of the efficient estimator sequence $\hat{\theta}_n$, the set $S(Z)$ denotes its support, and the function ϕ_0' denotes the directional derivative of ϕ at θ_0. This risk lower bound holds for all symmetric "bowl-shaped" loss functions, and parallels the familiar notion of the variance lower bound, establishing a sharp limit on the quality of estimation of the parameter $\phi(\theta_0)$ under directional differentiability. Second, I show that the estimator in (1.1) with adjustment terms $\hat{v}_{1,n}, \hat{v}_{2,n}$ solving a suitable sample analog of (1.3) attains the risk lower bound. This optimization problem takes a min-max form with a non-convex-concave objective function and, in general, can be computationally demanding. I discuss computational heuristics that help speed up the optimization and in some cases provide approximate closed-form solutions.

The finite-sample performance of the proposed estimator is investigated in a simulation study. I consider a simple setting, similar to Manski and Pepper (2000), in which the identified set for some real-valued parameter of interest is given by $[\max_{j\leqslant d_1}(\theta_{1,j}),\ \min_{k\leqslant d_2}(\theta_{2,k})]$, where $(\theta_1,\theta_2) = (\mathbb{E}_P(X_1),\mathbb{E}_P(X_2)) \in \mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$ for observable random vectors (X_1,X_2). Letting $(\bar{X}_{1,n},\bar{X}_{2,n})$ denote the corresponding sample means, one can estimate the bounds by $[\max_{j\leqslant d_1}(\bar{X}_{1,j,n}), \min_{k\leqslant d_2}(\bar{X}_{2,k,n})]$. However, the resulting estimates are generally biased towards each other, and, in practice, may be significantly tighter than the population bounds, potentially leading to erroneous conclusions. Therefore, it is customary to use bias-correction methods in practice (Kreider and Pepper, 2007; Chernozhukov et al., 2013). By extensive simulations, I compare the performance of the proposed estimator with the simple "plug-in" estimator and the existing bias-correction methods near the values of (θ_1,θ_2) where

the finite-sample bias is most problematic. These are the values (θ_1, θ_2) where the maximum/minimum are attained by multiple coordinates of θ_1 and θ_2 respectively,[4] so that the maximum/minimum functions are not fully differentiable. With the squared loss function, I find that the proposed estimator mildly reduces the bias but avoids substantial fluctuations in variance, compared to the alternatives.

As an application, I revisit the model of English auctions from Haile and Tamer (2003). In a setting with independent private values, the main primitive object of interest for the empirical analysis is the marginal distribution of valuations. The knowledge of this distribution allows one to forecast expected revenue and bidders surplus and study the effects of a change in the auction design. Under natural assumptions on bidders behavior, Haile and Tamer (2003) derived informative bounds on the distribution of valuations that take the form of point-wise minima and maxima of smooth transformations of the observed distribution of bids. I apply the methodology developed in this paper to construct estimators for the bounds on the distribution of valuations and the implied bounds on the optimal reserve price. Empirically calibrated simulations show that the resulting estimates are substantially sharper than the previously available ones.

This paper contributes to the literature on asymptotically efficient estimation in Econometrics and Statistics (e.g., Chamberlain, 1987, 1992; Newey, 1990, 1994a; Brown and Newey, 1998; Ai and Chen, 2003, 2012; Ackerberg et al., 2014; Kaido and Santos, 2014; Ibragimov and Hasḿinskii, 1981; Bickel et al., 1993; van der Vaart and Wellner, 1996; van der Vaart, 1988, 2000, and others). It is well-known that if $\hat{\theta}_n$ is asymptotically efficient for θ_0, and ϕ is fully (Hadamard) differentiable, the "plug-in" estimator $\phi(\hat{\theta}_n)$ is asymptotically efficient for $\phi(\theta_0)$ (e.g., van der Vaart, 1988). In this paper, I study asymptotically efficient estimators for $\phi(\theta_0)$ assuming only directional differentiability of ϕ, which allows to handle a new and

[4]Suppose that $\theta_{2,1}$ is the minimal component of θ_2 and it is well-separated from the rest, relative to the sampling uncertainty. Then, $\min_{k \leqslant d_2}(\bar{X}_{2,k,n}) = \bar{X}_{2,1,n}$ with probability close to one so that the plug-in estimator is approximately unbiased. On the other hand, if the minimal components of θ_2 are close to each other, the "plug-in" estimator is more likely to pick up the estimation errors in the components of $\bar{X}_{2,n}$. Similar intuition holds for the maximum function and the lower bound.

important class of parameters. Restricting attention to the class of estimators given in (1.1) allows to keep the analysis tractable and consistent with the literature. Efficient estimation under directional differentiability is also considered in Song (2014) and Fang (2018). Both papers aim to derive risk lower bounds for classes of estimators more general than (1.1), but the provided arguments turned out to be problematic.[5] That being said, very little is known about optimal estimation of non-differentiable functions. First, Blumenthal and Cohen (1968) show that in the experiment $\{Z \sim N(\theta, \sigma^2) : \theta \in \mathbb{R}\}$, $|Z|$ is a minimax for $|\theta|$, and in the experiment $\{(Z_1, Z_2) \sim N(\theta_1, \sigma^2) \times N(\theta_2, \sigma^2) : \theta_1, \theta_2 \in \mathbb{R}\}$, $\max(Z_1, Z_2)$ is a minimax esitmator for $\max(\theta_1, \theta_2)$. Second, Cai and Low (2011) derive a rate-optimal estimator (as $p \to \infty$) for a non-smooth functional $\phi(\theta) = p^{-1} \sum_{j=1}^{p} |\theta_j|$ in the Gaussian model $\{Z \sim N(\theta, I) : \theta \in \mathbb{R}^p\}$ based on suitable polynomial approximations. This indicates that the problem at hands is very hard and may not have a unique general solution. At the same time, for a class of estimators in (1.1) a complete and general answer can be provided. Another closely-related paper is Fang and Santos (2019). My work is complementary to theirs: I focus on efficient estimation, whereas they focus on valid inference in settings with directionally differentiable functions.

The rest of the paper is organized as follows. Section 1.2 provides the general setup and motivating examples and discusses the appropriate notion of directional differentiability. Section 1.3 elaborates on the optimality criterion, provides some background, and formulates the basic assumptions. Sections 1.4 and 1.5 establish the general risk lower bound under directional differentiability and construct efficient estimators. Section 1.6 presents a simulation study. Section 1.7 contains an empirical application. Section 1.8 discusses extensions, and Section 1.9 concludes.

[5] A more detailed discussion is provided in Section 1.4

1.2 Directionally Differentiable Parameters

1.2.1 General Setup

The main parameter of interest in this paper is $\phi(\theta_0)$, where θ_0 is an unknown but estimable feature of the distribution of the data, and ϕ is a known directionally differentiable function. In order to accommodate applications such as incomplete auction models or regression models with interval-valued data, I allow both θ_0 and $\phi(\theta_0)$ to take values in possibly infinite dimensional spaces. Specifically, I assume that $\theta_0 \in \mathbb{B}$ and $\phi : \mathbb{B} \to \mathbb{D}$ where $(\mathbb{B}, ||\cdot||_{\mathbb{B}})$ and $(\mathbb{D}, ||\cdot||_{\mathbb{D}})$ are Banach spaces. This includes $\mathbb{B} = \mathbb{R}^{d_\theta}$ and $\mathbb{D} = \mathbb{R}^{d_\phi}$ with the standard Euclidean norm as a special case.

Throughout the paper, I assume that the data $X_1^n \equiv (X_1, \ldots, X_n)$ are an i.i.d. sample drawn from a distribution $P \in \mathbf{P}$ of a random vector $X \in \mathbf{X}$.[6] Here, $\mathbf{P}$ denotes the model, i.e. the set of probability distributions (on a measurable space $(\mathbf{X}, \mathcal{B})$) that are plausible under the maintained assumptions. The set $\mathbf{P}$ may be explicitly indexed by finite- or infinite-dimensional parameters. The unknown parameter θ_0 takes value $\theta(P)$ when the distribution of the data is $P \in \mathbf{P}$.

Generic estimators for θ_0 and $\phi(\theta_0)$ are denoted by $\hat{\theta}_n : X_1^n \to \mathbb{B}$ and $\hat{\phi}_n : X_1^n \to \mathbb{D}$ respectively. The distributional convergence is understood in the Hoffman-Jørgensen sense (van der Vaart and Wellner, 1996), which does not require $\hat{\theta}_n$ and $\hat{\phi}_n$ to be measurable for each n. This fact is hidden from the notation throughout the text but highlighted in the Appendix when necessary. The distributional convergence denoted by $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow_{P_n} \mathbb{G}$ and $\sqrt{n}(\hat{\phi}_n - \phi_0) \rightsquigarrow_{P_n} \mathbb{W}$ is understood to be in $\mathbb{B}$ and in $\mathbb{D}$ respectively, with respect to the joint law $\prod_{i=1}^n P_n$ of X_1^n. The individual laws P_n may change with n.

The transpose of any vector a is denoted by a^T. The indicator functions are denoted by

[6] The i.i.d. setup is not essential: the asymptotic analysis relies on the notion of Local Asymptotic Normality which extends to non-i.i.d. settings via the limits of experiments framework. See Ibragimov and Hasḿinskii 1981; Le Cam 1986; van der Vaart 2000; Fang 2018.

$\mathbf{1}(S)$, which is equal to one if the statement S holds and to zero otherwise. For any pair of probability measures P and Q defined on the same measurable space, the ratio dP/dQ denotes the Radon-Nikodym derivative of the absolutely continuous part of P with respect to Q. For any sequences of constants a_n and b_n and random variables A_n and B_n, the symbol $A_n = o_{P_n}(a_n)$ means that A_n/a_n converges in probability to zero under P_n, and $B_n = O_{P_n}(b_n)$ means that B_n/b_n is bounded in probability under P_n.

1.2.2 Motivating Examples

Next, I present several motivating examples, some of which I revisit throughout the paper to fix ideas. These examples cover both finite and infinite-dimensional parameters and include models of treatment effects (Example 1.1), discrete choice (Example 1.2), English auctions (Example 1.3), regression models with interval-valued data (Example 1.4), and shape restrictions via projection (Example 1.5). To focus on the main ideas, the examples are simplified.

The first example, due to Manski and Pepper (2000, 2009), concerns estimation of bounds on average treatment effects.

Example 1.1 (Bounds on Average Treatment Effects)**.** Consider the standard potential outcomes framework. Let $D \in \{0, 1\}$ denote the treatment indicator, $Y(d) \in [\underline{y}, \overline{y}]$ denote the potential outcome under treatment $d \in \{0, 1\}$, $Y = DY(1)+(1-D)Y(0)$ denote the observed outcome, and $X \in \{x_1, \ldots, x_M\}$ denote an observed discrete covariate. The basic parameter of interest is $\mathbb{E}(Y(d)|X = x_m)$, i.e., the expected potential outcome under treatment d for a subpopulation with $X = x_m$. This parameter can only be point-identified under the assumption that the potential outcomes $(Y(0), Y(1))$ are statistically independent from D conditional on X, which may be hard to support in non-experimental settings. To provide a viable alternative, Manski and Pepper (2000) propose a number of weaker assumptions that deliver informative bounds on the parameter of interest, including the following Monotone

Instrumental Variables assumption. Suppose there is an order $x_1 \preccurlyeq \cdots \preccurlyeq x_M$ such that $x_j \preccurlyeq x_{j+1}$ implies

$$\mathbb{E}(Y(d)|X = x_j) \leqslant \mathbb{E}(Y(d)|X = x_{j+1})$$

for $d \in \{0, 1\}$ and all $j = 1, \ldots, M-1$. For example, letting Y denote wage, D indicate attending college, and X contain some measure of ability, it is reasonable to assume that the individuals with higher ability ($X = x_{j+1}$) are, on average, better off than their less talented peers ($X = x_j$) both in and out of college. Under this assumption, Manski and Pepper (2000) show that

$$\max_{j \leqslant m} \theta_{jd}(\underline{y}) \;\leqslant\; \mathbb{E}(Y(d)|X = x_m) \;\leqslant\; \min_{j \geqslant m} \theta_{jd}(\overline{y}),$$

where, for $y \in \{\underline{y}, \overline{y}\}$, $d \in \{0, 1\}$, and $j = 1, \ldots, M$,

$$\theta_{jd}(y) = \mathbb{E}(Y|X = x_j, D = d)P(D = d|X = x_j) + y \cdot P(D \neq d|X = x_j).$$

The above bounds on the expected potential outcomes can be used to obtain bounds on the average treatment effects, or strengthened under further monotonicity restrictions. Using similar ideas, Blundell et al. (2007) study changes in the distribution of wages accounting for selection into labor force, and Kreider et al. (2012) study the effects on food stamps on child health outcomes accounting for endogenous or misreported participation. See Ho and Rosen (2015) for a detailed review of recent applications. In this example, $\theta = (\theta_1, \theta_2) \in \mathbb{R}^m \times \mathbb{R}^{M-m+1}$ where $\theta_1 = (\theta_{jd}(\underline{y}))_{j=1}^m$ and $\theta_2 = (\theta_{jd}(\overline{y}))_{j=m}^M$, and the function $\phi : \mathbb{R}^{M+1} \to \mathbb{R}^2$ is given by

$$\phi(\theta) = \begin{pmatrix} \max_{j \leqslant m}(\theta_{1,j}) \\ \min_{k \leqslant M-m+1}(\theta_{2,k}) \end{pmatrix}.$$

This function is not fully differentiable at θ_0 if the maximum or the minimum are attained by multiple components of the corresponding subvector of θ_0. ■

The next example, due to Pakes et al. (2007, 2015) and Pakes (2010), concerns bounds on a real-valued parameter of interest in a partially-identified discrete-choice model.

Example 1.2 (Counterfactuals in Moment Inequality Models)**.** Suppose an agent chooses $y \in \mathbb{R}^{d_Y}$ from a set $\mathcal{Y} = \{y_1, \ldots, y_M\}$ to maximize her expected payoff $\mathbb{E}(\pi(y, Z, \gamma_0)|\mathcal{F})$, where Z is a vector of payoff-relevant variables, γ_0 is a vector of payoff parameters, and $\mathcal{F}$ is the agent's information set. Let Y^* denote the optimal choice, and assume that (Y^*, Z) are observed by the econometrician. Then, optimality of Y^* implies that for all $y' \in \mathcal{Y}$,

$$\mathbb{E}(\pi(y', Z, \gamma_0) - \pi(Y^*, Z, \gamma_0)|\mathcal{F}) \leqslant 0. \tag{1.4}$$

A common payoff specification is $\pi(y, Z, \gamma_0) = u(y, Z) + y^T\gamma_0$, where u is a known function (e.g., Pakes, 2010). Under suitable assumptions, the optimality condition in (1.4) implies that γ_0 must satisfy, for any $y, y' \in \mathcal{Y}$,

$$\mathbb{E}\left(\left(u(y', Z) - u(y, Z) + (y' - y)^T\gamma_0\right)\mathbf{1}(Y^* = y)\right) \leqslant 0$$

Therefore, the identified set for the vector of structural parameters $\gamma_0 \in \mathbb{R}^d$ is a convex polytope and it can be expressed as

$$\Gamma_0 = \{\gamma \in \mathbb{R}^{d_\gamma} : \mathbb{E}(m_{1j}(X) + m_{2j}(X)^T\gamma) \leqslant 0,\, j = 1, \ldots, J\}, \tag{1.5}$$

where m_{1j}, m_{2j} are known functions, and X is directly observed by the econometrician. Let $f(\gamma_0) = a + b^T\gamma_0$ denote a counterfactual of interest, representing, for instance, an expected change in profit. Assuming that Γ_0 is compact, the identified set for $f(\gamma_0)$ is given by $[L(\theta_0), U(\theta_0)]$ defined as

$$\begin{aligned} L(\theta_0) &= \min_{\gamma \in \mathbb{R}^{d_\gamma}} \{f(\gamma) \mid F(\theta_0, \gamma) \leqslant 0\}, \\ U(\theta_0) &= \max_{\gamma \in \mathbb{R}^{d_\gamma}} \{f(\gamma) \mid F(\theta_0, \gamma) \leqslant 0\}, \end{aligned}$$

where $\theta_0 \in \mathbb{R}^{2J}$ is a vector of moments containing $\mathbb{E}(m_{1j}(X))$ and $\mathbb{E}(m_{2j}(X))$ for all $j = 1, \ldots, J$ and the function $F(\theta_0, \gamma)$ defines the inequalities. In this example, $\mathbb{B} = \mathbb{R}^{2J}$, $\mathbb{D} = \mathbb{R}^2$, and the function $\phi : \mathbb{R}^{2J} \to \mathbb{R}^2$ is given by $\phi(\theta) = [L(\theta), U(\theta)]$. This function is not fully differentiable whenever the above optimization problems have multiple solutions. A conceptually different approach to identification in an overlapping class of models has been

developed in Galichon and Henry (2011) and Beresteanu et al. (2011), who characterize sharp identified sets for the structural parameters using tools from the theory of random sets. In particular, the so-called Artstein inequalities (Artstein, 1983) naturally fit the framework of the present paper. A detailed discussion of this matter, and the treatment of general moment inequality models, is provided in the Appendix. ■

The next example, due to Haile and Tamer (2003), concerns bounds on the distribution of valuations in English auctions.

Example 1.3 (English Auctions)**.** Consider a symmetric ascending auction with independent private values. Each bidder draws her valuation $V_i \in [\underline{v}, \overline{v}]$, independently of the others, from a distribution with a cumulative distribution function (CDF) denoted by F. Let B_i denote the final bid of player i. For simplicity, suppose that each auction has N bidders, and the reserve price is below $\underline{v}$. The main parameter of interest in the empirical analysis in this setting is the CDF of valuations F. To relate the unobserved valuations with the observed bids, Haile and Tamer (2003) assume that each player: (i) does not bid above her valuation and (ii) does not let the others win at a price she is willing to pay. Assumption (i) can be used to obtain an upper bound on the distribution of valuations

$$F(v) \leqslant \min_{i \leqslant N} \psi_i(G_{i:N}(v)),$$

where $G_{i:N}$ is the CDF of the i-th smallest bid, and $\psi_i : [0,1] \to [0,1]$ is a strictly increasing differentiable function.[7] In turn, Assumption (ii) can be used to obtain a lower bound using the distribution of the winning bid. Let $D([\underline{v}, \overline{v}], [0,1])$ denote the set of all cádlág functions from $[\underline{v}, \overline{v}]$ to $[0,1]$ (i.e., functions that are continuous from the right and have left limits evewyehrer) endowed with the supremum norm. Focusing on the upper bound presented above, in this example, $\mathbb{B} = D([\underline{v}, \overline{v}], [0,1])^N$, $\mathbb{D} = D([\underline{v}, \overline{v}], [0,1])$,

[7] This function relates the marginal distribution of the order statistics of i.i.d. random variables with the parent distribution. More details are provided in Section 1.7.

$\theta_0 = (\psi_1(G_{1:N}), \ldots, \psi_N(G_{N:N})) \in \mathbb{B}$ and $\phi : \mathbb{B} \to \mathbb{D}$, is defined by

$$\phi(\theta)(v) = \min_{i \leqslant N}(\theta_{0,i}(v)).$$

This function is not fully differentiable if the minimum is attained by multiple $\theta_{0,i}$ for at least one $v \in [\underline{v}, \overline{v}]$. For example, if the bids are i.i.d., all $\psi_i(G_{i:N}(v))$ will coincide for all $v \in [\underline{v}, \overline{v}]$. The bounds on the distribution of valuations can be translated into the bounds on the expected revenue, bidders surplus, and optimal reserve price; see Haile and Tamer (2003). In the same setting, Chesher and Rosen (2017) characterize the sharp bounds on the distribution of valuations using tools from the theory of random sets. Aradillas-López et al. (2013a) provide bounds on the expected revenue and bidders surplus in auctions with correlated private values. ■

The next example, due to Beresteanu and Molinari (2008) and Bontemps et al. (2012), deals with a regression model with interval-valued outcomes.

Example 1.4 (Interval Outcome Regression)**.** Let $Y \in \mathbb{R}$ be an outcome variable, $Z \in \mathbb{R}^{d_Z}$ be a vector of covariates, and $\beta_0 \in \mathbb{R}^{d_Z}$ be a vector of coefficients for the best linear prediction

$$Y = Z^T\beta_0 + \varepsilon, \qquad \mathbb{E}(\varepsilon Z) = 0.$$

Assume that $Y_L \leqslant Y \leqslant Y_U$ almost surely and the researcher only observes (Z, Y_L, Y_U). One parameter of interest is $\gamma_0 = p^T\beta_0$, with known $p \in \mathbb{R}^{d_Z}$, representing, for example, a coordinate projection. Bontemps et al. (2012) derived the closed-form expressions for the bounds on γ_0, given by

$$\inf_{\beta \in B_0} p^T\beta = \mathbb{E}(b_0^T Z Y_L + \min\{b_0^T Z, 0\}(Y_U - Y_L)),$$
$$\sup_{\beta \in B_0} p^T\beta = \mathbb{E}(b_0^T Z Y_L + \max\{b_0^T Z, 0\}(Y_U - Y_L)),$$

where $b_0 = (\mathbb{E}(ZZ^T))^{-1}p \in \mathbb{R}^{d_Z}$, and B_0 is the sharp identified set for β_0. Denote $\theta_0 = (\psi_0, b_0)$, where $\psi_0 : \mathbb{R}^{d_Z} \to \mathbb{R}^2$ is given by

$$\psi_0(b) = \begin{pmatrix} \mathbb{E}(b^T Z Y_L + \max\{b^T Z, 0\}(Y_U - Y_L)) \\ \mathbb{E}(b^T Z Y_L + \min\{b^T Z, 0\}(Y_U - Y_L)) \end{pmatrix}.$$

Letting $l^\infty(T)$ denote the set of all bounded real-valued functions defined on T endowed with the supremum norm, it is convenient to view $\psi_0 \in l^\infty(B)$ for some compact set B containing b_0 in its interior. Then, $\mathbb{B} = l^\infty(B) \times \mathbb{R}^{d_Z}$, $\mathbb{D} = \mathbb{R}^2$ and $\phi : \mathbb{B} \to \mathbb{D}$ is defined by $\phi(\theta) = \psi(b)$ for any $(\psi, b) \in \mathbb{B}$. This function is not fully differentiable if $P(b_0^T Z = 0) > 0$. More generally, one can consider any parameter of the form $\psi(\beta)$, where both β and ψ are unknown, but root-n estimable, and ψ is potentially only directionally differentiable. For example, forecasts in regression kink models share a similar structure; see Hansen (2017). ■

The final example concerns quantile regression models. Due to the potential misspecification, the quantile regression function may not be monotone, which complicates interpretation (Bassett and Koenker, 1982; Angrist et al., 2006). To avoid this problem, Fang (2018) proposes projecting the curve onto a suitable set of monotone functions.[8]

Example 1.5 (Quantiles without Crossing)**.** Let $Y \in \mathbb{R}$ and $Z \in \mathbb{R}^d$ denote the outcome variable and the set of covariates correspondingly, and consider the quantile regression model

$$\beta(\tau) = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \, \mathbb{E}(\rho_\tau(Y - Z^T\beta)),$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}\{u \leqslant 0\})$. Denote the quantile regression process, for a fixed value of z, by $\theta(\tau) = z^T\beta(\tau)$. Let $\mathcal{T} = [\varepsilon, 1 - \varepsilon]$ with $\varepsilon \in (0, 1/2)$, and view $\theta : \mathcal{T} \to \mathbb{R}$ as an element of $L^2(\mathcal{T})$, denoting the space of square-integrable functions with respect to the Lebesgue measure. To impose monotonicity, one may project $\theta(\tau)$ onto the set $\Lambda \subset L^2(\mathcal{T})$ of all monotonically increasing functions:

$$\phi(\theta) = \Pi_\Lambda \theta \equiv \underset{\lambda \in \Lambda}{\operatorname{argmin}} \, ||\theta_0 - \lambda||_{L_2(\mathcal{T})}.$$

Since Λ is a convex cone, the projection exists and is unique. In this example, $\mathbb{B} = L^2(\mathcal{T})$, $\mathbb{D} = \Lambda$, and $\phi : L^2(\mathcal{T}) \to \Lambda$ is defined by $\phi(\theta) = \Pi_\Lambda \theta$. The projection map is not fully differentiable at all points that are projected on a vertex of Λ. ■

[8] This provides an alternative to the monotone rearrangement operator of Chernozhukov et al. (2010).

1.2.3 Hadamard Directional Differentiability

In the above examples, there exist points θ_0 at which the corresponding function ϕ is not fully differentiable. However, at such points, ϕ remains directionally differentiable in the following sense:

Definition 1.1. *A function $\phi : \mathbb{B} \to \mathbb{D}$ is Hadamard directionally differentiable at θ_0 if there is a continuous function $\phi'_0 : \mathbb{B} \to \mathbb{D}$ such that, for any $h_n \to h$ in $\mathbb{B}$, and any $t_n \downarrow 0$,*

$$\lim_{n\to\infty} \left|\left| \frac{\phi(\theta_0 + t_n h_n) - \phi(\theta_0)}{t_n} - \phi'_0(h) \right|\right|_{\mathbb{D}} = 0. \tag{1.6}$$

If the above holds for each $h \in \mathbb{B}_0 \subset \mathbb{B}$, it is said that ϕ is directionally differentiable at θ_0 tangentially to $\mathbb{B}_0$. In this case, the domain of ϕ'_0 is $\mathbb{B}_0$.

Intuitively, a function is directionally differentiable at θ_0 if it can be linearly approximated in each direction around θ_0, and the approximation is suitably continuous. To compare, a function ϕ is *Hadamard fully differentiable* if the derivative ϕ'_0, satisfying (1.6), is a continuous *linear* function. That is, full differentiability implies directional differentiability, and the only distinction between the two notions is the potential non-linearity of the directional derivative (Shapiro, 1990).

In this paper, in addition to Hadarmard directional differentiability of the function ϕ, I require that the directional derivative be Lipchitz-continuous.

Assumption 1.2.1 (Restrictions on ϕ)**.** *The map $\phi : \mathbb{B} \to \mathbb{D}$ is directionally Hadamard differentiable at θ_0 tangentially to $\mathbb{B}_0$, as in Definition 1.1. Moreover, the directional derivative $\phi'_0 : \mathbb{B}_0 \to \mathbb{D}$ is Lipchitz-continuous. That is,*

$$||\phi'_0(x) - \phi'_0(y)||_{\mathbb{D}} \leqslant C_{\phi'} ||x - y||_{\mathbb{B}}$$

for all $x, y \in \mathbb{B}_0$, for some $C_{\phi'} < \infty$.

Since continuous linear functions are Lipchitz-continuous, this assumption is satisfied whenever ϕ is fully differentiable. Otherwise, it only imposes a mild restriction: the directional derivative is a "partially linear" function with different "slopes" in different regions

of the domain, so the assumption merely rules out unbounded "slopes". Moreover, in most applications, the function ϕ itself is Lipschitz-continuous, in which case Assumption 1.2.1 is automatically satisfied; see Shapiro (1990).

1.2.3.1 Examples Revisited

To fix ideas, I will focus on Examples 1 and 3 throughout the paper. The remaining examples are discussed in the Appendix.

Example 1 (Continued)**.** Focus on the upper bound $\phi(\theta_0) = \min_{j \leqslant d}(\theta_{0,j})$ with $\theta_0 \in \mathbb{R}^d$. For each $h = (h_1, \ldots, h_d)^T$, the directional derivative is equal to

$$\phi_0'(h) = \min_{j \in B(\theta_0)}(h_j), \tag{1.7}$$

where $B(\theta_0) = \{j : \theta_{0,j} = \min_i(\theta_{0,i})\}$ is the set of indices of the components of θ_0 that attain the minimum. That is, the function ϕ is fully differentiable at θ_0 if there is a unique minimal component, and only directionally differentiable otherwise. The directional derivative satisfies Assumption 1.2.1 with $C_{\phi'} = 1$. Similar arguments hold for the lower bound, and for both bounds simultaneously. ■

Example 3 (Continued)**.** Assume that $N = 2$, so that $\theta_0 \in D([\underline{v}, \overline{v}], [0,1])^2$ is given by $\theta_0(v) = (\theta_{1,0}(v), \theta_{2,0}(v)) = (\psi_1(G_{1:2}(v)), \psi_2(G_{2:2}(v)))$. Recall that $\phi(\theta)(v) = \min\{\theta_1(v), \theta_2(v)\}$, and define the sets

$$\begin{aligned} S_1(\theta_0) &= \{v \in [\underline{v}, \overline{v}] : \theta_{1,0}(v) < \theta_{2,0}(v)\} \\ S_2(\theta_0) &= \{v \in [\underline{v}, \overline{v}] : \theta_{2,0}(v) < \theta_{1,0}(v)\} \\ S_0(\theta_0) &= \{v \in [\underline{v}, \overline{v}] : \theta_{1,0}(v) = \theta_{2,0}(v)\}. \end{aligned} \tag{1.8}$$

The directional derivative $\phi_0' : D([\underline{v}, \overline{v}], [0,1])^2 \to D([\underline{v}, \overline{v}], [0,1])$ is given by

$$\begin{aligned} \phi_0'(h)(v) = h_1(v) \cdot \mathbf{1}(v \in S_1(\theta_0)) + h_2(v) \cdot \mathbf{1}(v \in S_2(\theta_0)) \\ + \min\{h_1(v), h_2(v)\} \cdot \mathbf{1}(v \in S_0(\theta_0)) \end{aligned} \tag{1.9}$$

for any $h = (h_1, h_2) \in D([\underline{v}, \overline{v}], [0,1])^2$. Therefore, whenever the set S_0 is non-empty, the function ϕ is only directionally differentiable. For example, if the bids are i.i.d., then $\psi_i(G_{i:2}) = G$ for $i = 1, 2$, so that $S_0 = [\underline{v}, \overline{v}]$. The directional derivative satisfies Assumption 1.2.1 with $C'_\phi = 1$. ■

1.3 Local Asymptotic Minimaxity

This section formally defines the efficiency criterion and formulates the basic assumptions of the paper. Before diving into the technical details, I discuss the general idea of the criterion.

1.3.1 General Idea

Intuitively, a "good" estimator should not deviate from the estimand too much, too often. The notion of risk provides a way to quantify this intuition. To elaborate, recall that the data $X_1, \ldots, X_n$ are an i.i.d. sample with a common distribution $P \in \mathbf{P}$, where $\mathbf{P}$ denotes the model, and the parameter θ_0 takes value $\theta(P)$ when the underlying distribution is $P \in \mathbf{P}$. Let $\hat{\phi}_n$ denote a generic root-n consistent estimator for the target parameter $\phi(\theta_0)$. Let l denote a non-negative "bowl-shaped" loss function, which specifies penalties, $l(\sqrt{n}(\hat{\phi}_n - \phi(\theta_0)))$, imposed when the estimator deviates from the estimand. Then, the risk of the estimator $\hat{\phi}_n$ under the distribution P is defined as $\mathbb{E}_P(l(\sqrt{n}(\hat{\phi}_n - \phi(\theta_0))))$. For a given loss function and fixed n, it is understood that the smaller the risk, the better the estimator.

Additionally, since the distribution P is *ex ante* unknown, beyond the assumption that $P \in \mathbf{P}$, a good estimator should perform well in some overall sense within $\mathbf{P}$. For example, one may take the Bayesian approach and construct estimators that minimize the average risk, calculated over some prior belief about $\mathbf{P}$, or the minimax approach and construct estimators that minimize the worst-case risk within $\mathbf{P}$ (see, e.g., Lehmann and Casella (2006) for the discussion of these and other approaches). However, one often lacks prior knowledge about the relative likelihood of the plausible distributions (especially, in semi- and non-parametric

models), while tailoring the estimator to the least favorable distribution may worsen its performance at other, potentially more empirically relevant distributions.

To gain tractability, one may take a more local approach. As the sample size increases, the true distribution P of the observed data can be better located within the model $\mathbf{P}$. Therefore, one may focus on the appropriate "local neighborhoods" $V_n(P) \subset \mathbf{P}$ around P and evaluate different estimators by their asymptotic worst-case risk within such neighborhoods. This line of thought leads to the notion of Local Asymptotic Minimaxity. Formally, an estimator sequence $\{\hat{\phi}_n\}$ is Locally Asymptotically Minimax (LAM) if it minimizes the asymptotic locally-worst-case risk, that is,

$$\liminf_{n\to\infty} \sup_{\tilde{P}\in V_n(P)} \mathbb{E}_{\tilde{P}} \left(l \left(\sqrt{n}(\hat{\phi}_n - \phi(\theta(\tilde{P}))) \right) \right). \tag{1.10}$$

The local neighborhoods $V_n(P)$ shrink to P as n approaches infinity and only contain distributions that are hard to distinguish from P empirically. The discussion below makes this definition rigorous, providing the necessary background, stating the main assumptions, and discussing the choice of the local neighborhoods and loss functions.

1.3.2 Background and Assumptions

I start by defining the main components of the local asymptotic framework, following the literature on semiparametric efficiency (e.g., Bickel et al., 1993). The following notation is used recurrently. For a probability measure P on $(\mathbf{X}, \mathcal{B})$, the spaces $L_2(P)$ and $L_2^0(P)$ are defined as

$$L_2(P) = \left\{ h : \mathbf{X} \to \mathbb{R} \;\middle|\; \int h^2 dP < \infty \right\},$$

$$L_2^0(P) = \left\{ h : \mathbf{X} \to \mathbb{R} \;\middle|\; \int h^2 dP < \infty, \int h dP = 0 \right\}.$$

These spaces are endowed with the standard $L_2(P)$ norm $||h||_{2,P} = (\int h^2 dP)^{1/2}$ and scalar product $\langle h_1, h_2 \rangle_P = \int h_1 h_2 dP$. For any subset H, of $L_2(P)$, $\bar{H}$ denotes its closure with

respect to $||\cdot||_{2,P}$. To simplify exposition, I assume that the model $\mathbf{P}$ is dominated by a positive, sigma-finite measure μ on $(\mathbf{X}, \mathcal{B})$.

1.3.2.1 Smooth Parametric Submodels and Tangent Sets

The idea of local asymptotic analysis is to study the behavior of the parameters and estimators of interest along suitable submodels of $\mathbf{P}$ passing through P. Following the literature, I consider smooth parametric sumbodels and scores defined as follows.

Definition 1.2 (Smooth Parametric Submodels and Scores)**.** *A smooth parametric submodel $t \mapsto P_{t,h}$ is a mapping defined on $[0, \varepsilon)$ for some $\varepsilon > 0$, such that (i) $P_{t,h}$ is a probability distribution for each t; (ii) $P_{0,h} = P$; and (iii) for some measurable function $h : \mathbf{X} \to \mathbb{R}$,*

$$\int \left(\frac{\sqrt{p_{t,h}} - \sqrt{p}}{t} - \frac{1}{2}\sqrt{p}h \right)^2 d\mu \to 0 \quad \text{as } t \downarrow 0. \tag{1.11}$$

Such h is called the score for the submodel $\{P_{t,h}\}$. Here $p_{t,h} = dP_{t,h}/d\mu$ and $p = dP/d\mu$ denote the densities of $P_{t,h}$ and P with respect to μ.

The score h, defined above, is a quadratic-mean version of the familiar parametric score, defined by $\partial \log p_{t,h}(x)/\partial t|_{t=0}$. Any score h automatically satisfies $\mathbb{E}_P(h) = 0$ and $\mathbb{E}_P(h^2) < \infty$, so that $h \in L_2^0(P)$. The collection of all scores corresponding to the submodels $\{P_{t,h}\} \subset \mathbf{P}$ is called the tangent set.

Definition 1.3 (Tangent Set)**.** *The set of all scores corresponding to the submodels $\{P_{t,h}\} \subset \mathbf{P}$ is called the tangent set and denoted by*

$$T(P) = \{h \in L_0^2(P) \mid h \text{ satisfies (1.11) for some } \{P_{t,h}\} \subset \mathbf{P}\}. \tag{1.12}$$

The tangent set depends on both the distribution P and the model $\mathbf{P}$ and describes the informational content of the assumption $P \in \mathbf{P}$. It is directly related to both construction of efficient estimators (e.g., Bickel et al., 1993) and existence of specification tests with non-trivial power (Chen and Santos, 2018). Assumptions on $\mathbf{P}$ may translate into further

restrictions on the tangent set through the requirement $\{P_{t,h}\} \subset \mathbf{P}$. If $T(P) = L_2^0(P)$, the tangent set is said to be unrestricted; otherwise, it is restricted. In the latter case, the tangent set typically forms a linear subspace of $L_2^0(P)$, but in some cases $T(P)$ can be a convex cone, e.g., in some moment inequality models.[9]

Throughout the paper, I assume that the tangent set is a linear space, as recorded below. A partial extension of the main results to convex cones and some issues associated with such settings are discussed in Section 1.8.

Assumption 1.3.1 (Random Sampling and Restrictions on the Model)**.** *The researcher observes an i.i.d. sample* $\{X_i\}_{i=1}^n$ *of* $X \in \mathbf{X}$ *from* $P \in \mathbf{P}$*. The model* $\mathbf{P}$ *and the distribution* $P \in \mathbf{P}$ *are such that tangent set* $T(P)$ *is a linear subspace of* $L_0^2(P)$*.*

1.3.2.2 Differentiable Parameters and Regular Estimators

For a submodel $\{P_{t,h}\} \subset \mathbf{P}$ with a score $h \in T(P)$, denote $P_{n,h} \equiv P_{1/\sqrt{n},h}$. The parameter $\theta_0 = \theta(P)$ is assumed to be differentiable in the following sense.

Definition 1.4 (Path-Wise Differentiable Parameters)**.** *A parameter* $\theta(P) \in \mathbb{B}$ *is differentiable relative to a tangent set* $T(P)$ *if there is a continuous linear functional* $\theta_0' : \bar{T}(P) \to \mathbb{B}$*, such that*

$$\sqrt{n}(\theta(P_{n,h}) - \theta(P)) \;\to\; \theta_0'(h) \qquad \text{in } \mathbb{B}, \text{ as } n \to \infty.$$

The functional $\theta_0'(h)$ *is called the path-wise derivative of* $\theta(P)$*.*

Assumption 1.3.2 (Diferentiability of $\theta(P)$)**.** *The parameter* $\theta(P)$ *is differentiable relative to the tangent set* $T(P)$*, according to Definitions 1.2, 1.3, and 1.4.*

Path-wise differentiability guarantees existence of the estimators with nice asymptotic behavior. The path-wise derivative θ_0' is crucial in characterizing the asymptotic efficiency

[9]The tangent set $T(P)$ is a cone by construction. If $h \in L_2^0(P)$ corresponds to a submodel $\{P_t\}$ then $ah \in L_2^0(P)$ for any $a \geqslant 0$ corresponds to the submodel $\{P_{at}\}$. Therefore, $T(P)$ is a collection of rays i.e. a cone. For a detailed discussion, see van der Vaart (1988).

bound for $\theta(P)$, which is discussed in more details in Section 1.3.2.3.[10] With i.i.d. data, this assumption limits the analysis to parameters estimable at the root-n rate. Examples include moments, distribution functions, quantile functions, parametric components in semi-parametric models, and smooth functions of those. Differentiable parameters are typically estimated with regular estimators.

Definition 1.5 (Regular Estimator)**.** *A sequence of estimators $\hat{\theta}_n : X_1^n \to \mathbb{B}$ for a parameter $\theta(P) \in \mathbb{B}$ is regular, if*

$$\sqrt{n}(\hat{\theta}_n - \theta(P_{n,h})) \overset{P_{n,h}}{\rightsquigarrow} \mathbb{G} \qquad (\text{in } \mathbb{B})$$

for all $h \in T(P)$, where $\mathbb{G}$ is a tight random element in $\mathbb{B}$ that does not depend on h.

Regularity is a desirable property: A small disappearing perturbation of the distribution of the data should not affect the limit distribution of the estimator. For example, sample averages, empirical distribution and quantile functions, and smooth functions of those are regular estimators for the corresponding population parameters.

1.3.2.3 Convolution Theorem and Best Regular Estimators

The efficient estimator for $\phi(\theta_0)$, developed in the sequel, relies on the notion of the best regular estimator for θ_0, discussed below. Consider estimating a differentiable parameter $\theta_0 = \theta(P)$. The Convolution Theorem states that the asymptotic distribution of *any regular estimator* $\hat{\theta}_n$ can be represented as a convolution of a centered Gaussian random element $\mathbb{G}_0$ and an independent "noise term" $\mathbb{W}$, that is

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{P}{\rightsquigarrow} \mathbb{G}_0 + \mathbb{W}.$$

Since convolution increases variance, the "best possible" limit among regular estimators is $\mathbb{G}_0$, and its variance-covariance matrix of $\mathbb{G}_0$ is known as the *efficiency bound*. Any regular

[10] The concept of path-wise derivative originated in Koshevnik and Levit (1976) and Pfanzagl (1982) for Euclidean parameters and was extended to general normed spaces in van der Vaart (1988).

estimator that attains this limit is called the *best regular estimator*. The covariance structure and the support of $\mathbb{G}_0$ are determined by the path-wise derivative θ_0' and the tangent set $T(P)$ (see Theorems 1.3 and 1.4 in the Appendix).[11]

Next, consider estimating $\phi(\theta_0)$ with a fully Hadamard differentiable function ϕ with derivative ϕ_0' at θ_0. One can show that $\phi(\theta(P))$ is also a differentiable parameter, and the distributional limit of any regular estimator $\hat{\phi}_n$ satisfies

$$\sqrt{n}(\hat{\phi}_n - \phi(\theta_0)) \overset{P}{\rightsquigarrow} \phi_0'(\mathbb{G}_0) + \mathbb{W}',$$

where $\mathbb{G}_0$ is the same as in the previous display, and $\mathbb{W}'$ is an independent "noise term" (e.g., van der Vaart, 1988). In the same fashion as above, the best regular estimator sequence converges in distribution to $\phi_0'(\mathbb{G}_0)$, which is also a centered Gaussian random element, since the derivative ϕ_0' is linear. It follows from the Delta-method that if $\hat{\theta}_n$ is best regular for θ_0, the "plug-in" estimator $\phi(\hat{\theta}_n)$ is best regular for $\phi(\theta_0)$.

When estimating differentiable parameters, it is without loss of generality to focus on regular estimators, because best regular estimators are also asymptotically minimum-variance unbiased (when applicable) and locally asymptotically minimax among all estimators (e.g., van der Vaart, 2000). However, for parameters of the form $\phi(\theta_0)$ where ϕ is only directionally differentiable, regular and asymptotically unbiased estimators do not exist (van der Vaart, 1991; Hirano and Porter, 2012), so that it is necessary to consider larger classes of competing estimators.

1.3.3 LAM Risk and Directional Differentiability

Having introduced the notions of smooth parametric submodels and tangent sets, I am in position to define the optimality criterion rigorously. Following the literature, I define the

[11] For example, to construct the best regular estimator for $\theta_0 \in \mathbb{R}^d$, one has to find $\tilde{\theta}$ such that $\theta_0'(h) = \mathbb{E}_P(\tilde{\theta}h)$, project such $\tilde{\theta}$ onto $T(P)$, denoting the projection by ψ_θ, and seek an estimator such that $\sqrt{n}(\hat{\theta}_n - \theta_0) = n^{-1/2}\sum_{i=1}^n \psi_\theta(X_i) + o_P(1)$.

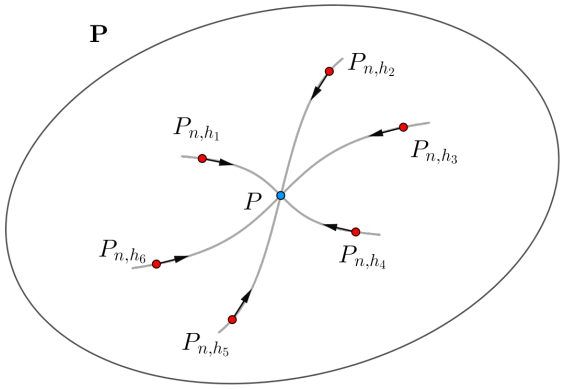


Figure 1.1: Example of a Local Neighborhood with $I = \{h_1, \ldots, h_6\}$.

LAM risk as[12]

$$\sup_{I \subset T(P)} \liminf_{n \to \infty} \sup_{h \in I} \mathbb{E}_{P_{n,h}} \left\{ l \left(\sqrt{n}(\hat{\phi}_n - \phi(\theta(P_{n,h}))) \right) \right\}, \tag{1.13}$$

where I denotes an arbitrary *finite* subset $I \subset T(P)$ of the tangent set, and $P_{n,h}$ denotes a probability distribution corresponding to a smooth parametric submodel $\{P_{t,h}\} \subset \mathbf{P}$ with a score $h \in T(P)$ with $t = 1/\sqrt{n}$. In the notation of Equation (1.10), the local neighborhoods are $V_n(P) = \{P_{n,h} : h \in I\}$. Figure 1.1 illustrates.

The restriction to finite neighborhoods is made for two reasons. First, when the local neighborhoods are too rich, the sharp lower bound for the local asymptotic maximum risk may be infinite (see van der Vaart, 1988). In such case, every estimator is "optimal", which makes the criterion meaningless. Second, to construct optimal estimators, one has to establish weak convergence uniformly over the local neighborhoods, which may be impossible if the neighborhoods are too large.

Next, I discuss the difficulties associated with deriving LAM estimators for parameters expressed via directionally differentiable functions. First, as in the previous section, consider estimating a differentiable parameter θ_0 with a regular estimator $\hat{\theta}_n$. Using a representation from the Convolution Theorem, one can argue that the LAM risk of any such estimator is

[12]See e.g., van der Vaart (1988); van der Vaart and Wellner (1996); Hirano and Porter (2009); Fang (2018).

$\mathbb{E}(l(\mathbb{G}_0 + \mathbb{W}))$. By Anderson's Lemma, it is larger than $\mathbb{E}(l(\mathbb{G}_0))$ for any symmetric quasi-convex loss function l. Moreover, this lower bound turns out to hold among all estimators, so the best regular estimator $\hat{\theta}_n$ whose distributional limit is $\mathbb{G}_0$ is also LAM. Note that in this argument, the random noise $\mathbb{W}$ in the expression for the LAM risk is replaced with a constant $w = 0$.

Now, consider estimating $\phi(\theta_0)$. Using a suitable Delta-Method, one can show that $\sqrt{n}(\phi(\hat{\theta}_n) - \phi(\theta_0))$ converges in distribution to $\phi_0'(\mathbb{G}_0 + \mathbb{W})$, for any regular $\hat{\theta}_n$, where ϕ_0' denotes the Hadamard directional derivative of ϕ at θ_0. If ϕ is fully differentiable, ϕ_0' is in fact linear, so $\phi_0'(\mathbb{G}_0 + \mathbb{W}) = \phi_0'(\mathbb{G}_0) + \phi_0'(\mathbb{W})$. Here, the first summand is also Gaussian, so the argument from the preceding paragraph still applies. However, if ϕ is only directionally differentiable, one can show that the LAM risk takes the form:

$$\sup_{h \in T(P)} \mathbb{E}(l(\phi_0'(\mathbb{G}_0 + \mathbb{W} + \theta_0'(h)) - \phi_0'(\theta_0'(h))). \tag{1.14}$$

Here, since ϕ_0' is now non-linear, the terms $\phi_0'(\theta_0'(h))$ do not cancel out. Recall from the previous section that $\mathbb{W}$ is a random noise whose distribution depends on $\hat{\theta}_n$. Song (2014) and Fang (2018) note that, in order to derive a practically useful lower bound, it would be desirable to replace $\mathbb{W}$ by a constant. In the absence of a general result in the spirit of Anderson's Lemma, it is a complicated task. To this end, Song (2014) and Fang (2018) suggest applying purification arguments from Dvoretzky et al. (1951) and Feinberg and Piunovskiy (2006) correspondingly. Essentially, these techniques allow to replace a randomized decision rule $\mathbb{W}_0$ with a deterministic rule $w(z)$, for an arbitrary finite number of loss functions $\rho_j(z, w) = l(\phi(z + w + \theta_0'(h_j)) - \phi(\theta_0'(h_j)))$. Note, however, that $w(z)$ is still a function of the state variable z, and cannot, in general, be replaced by a constant w. In special cases when it can, the resulting lower bound will take the form (1.3) and the optimal estimator will take the form (1.1). In general, however, other types of estimators may be optimal. For example, if l is convex, (1.14) is bounded from below by

$$\mathbb{E}_{\mathbb{G}_0}\left(l(\mathbb{E}_{\mathbb{W}}(\phi_0'(\mathbb{G}_0 + \mathbb{W} + \theta_0'(h)) - \phi'(\theta_0'(h))) \right),$$

by Jensen's inequality and independence of $\mathbb{G}_0$ and $\mathbb{W}$. This suggests an estimator of the form $\int \phi(\hat{\theta}_n + w/\sqrt{n})dF(w)$, which, intuitively, smooths out the non-differentiability of ϕ. Studying the properties of such estimators is beyond the scope of this paper and left for further research.

1.3.4 Loss Functions

An essential ingredient in the LAM-analysis is the loss function. It specifies which deviations of the estimator from the estimand should be punished relatively more than the others, and by how much. In practice, the loss function can be used to "fine-tune" the estimator (e.g. specify the relative importance of different dimensions of the target parameter, or focus on a subvector), address sensitivity to outliers in the data (e.g., consider the absolute loss instead of quadratic loss), or boost computation (e.g., pick a smooth or convex function). In theory, the loss function must ensure that the LAM risk is finite for at least one estimator, for otherwise the optimality criterion becomes meaningless (see, e.g., Lemma 3.1 in Fang, 2018).

Following the literature, I consider a large family of symmetric "bowl-shaped" loss functions, which are appropriate for most applications.

Assumption 1.3.3 (Loss Functions)**.** *The loss function $l : \mathbb{D} \to \mathbb{R}_+$ is sub-convex. That is, the lower level sets $\{x \in \mathbb{D} : l(x) \leqslant c\}$ are closed, convex and symmetric.*

Any sub-convex loss function must be lower semi-continuous and satisfy $l(-x) = l(x)$. This assumption rules out asymmetric loss functions, but allows, for example, for different weights along different dimensions of the argument, and for discontinuities. Some examples are provided below.

- For $x \in \mathbb{R}^d$, one can consider a weighted quadratic loss, absolute loss, or maximum

loss, with $w_1, \ldots, w_d \geqslant 0$:

$$\begin{aligned} l(x) &= w_1 x_1^2 + w_2 x_2^2 + \ldots + w_d x_d^2, \\ l(x) &= w_1 |x_1| + w_2 |x_2| + \ldots + w_d |x_d|, \\ l(x) &= \max\{w_1 |x_1|, w_2 |x_2|, \ldots, w_d |x_d|\}. \end{aligned}$$

Adjusting the weights allows to specify the relative importance of the coordinates.

- For $x \in l^\infty(S)$, one can consider the supremum loss or focus on a finite-dimensional slice, for some $s_1, \ldots, s_d \in S$ and $w_1, \ldots, w_d \geqslant 0$:

$$\begin{aligned} l(x) &= \sup_{s \in S} |x(s)|, \\ l(x) &= w_1 x(s_1)^2 + w_2 x(s_2)^2 + \cdots + w_d x(s_d)^2. \end{aligned}$$

- For $x \in L^2([a, b])$, one can consider a weighted L_2-loss, with bounded $w(t) \geqslant 0$,

$$l(x) = \int_a^b w(t) x^2(t) dt,$$

or focus on a finite-dimensional slice in the same fashion as above.

- In any of the above examples, one can consider a zero-one loss, defined as

$$l(x) = \mathbf{1}\{x \notin A\},$$

where A is a a closed convex set symmetric around the origin.

1.4 Risk Lower Bound

In this section, I formally derive the lower bound for LAM risk for all estimators of the form

$$\hat{\phi}_n = \phi\left(\hat{\theta}_n + \frac{\hat{v}_{1,n}}{\sqrt{n}}\right) + \frac{\hat{v}_{2,n}}{\sqrt{n}}, \tag{1.15}$$

where $\hat{v}_{1,n}, \hat{v}_{2,n}$ are adjustment terms converging in probability (under P) to some constants.

Theorem 1.1 below presents the general result, and Corollary 1.1.1 specializes to Euclidean parameters. To state the general result, some new notation is required. Recall that

the path-wise derivative is a continuous map $\theta_0' : \bar{T}(P) \to \mathbb{B}$. By the Riesz representation theorem, for any $b^* \in \mathbb{B}^*$ (the continuous dual of $\mathbb{B}$), there is an element $\tilde{\theta}_{b^*} \in \bar{T}(P)$ such that $b^*(\theta_0'(h)) = \langle \tilde{\theta}_{b^*}, h \rangle_{2,P}$ for all $h \in \bar{T}(P)$. Such $\tilde{\theta}_{b^*}$ is called the canonical gradient of θ_0' in direction b^*.

Theorem 1.1 (General Lower Bound)**.** *Let Assumptions 1.2.1, 1.3.1, 1.3.2, 1.3.3, and assume that the infimum in the display below can be attained. Then, for any estimator sequence of the form (1.15),*

$$\sup_{I \subset T(P)} \liminf_{n\to\infty} \sup_{h \in I} \mathbb{E}_{P_{n,h}} \left\{ l\left(\sqrt{n}(\hat{\phi}_n - \phi(\theta(P_{n,h}))) \right) \right\} \geqslant \inf_{(v_1,v_2)\in\mathbb{B}\times\mathbb{D}} \sup_{s \in S(\mathbb{G}_0)} \mathbb{E}\left\{ l\left(\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2 \right) \right\},$$

where I is an arbitrary finite subset of the tangent set $T(P)$, $\mathbb{G}_0$ denotes the distributional limit of the best regular estimator sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$, and $S(\mathbb{G}_0) \subset \mathbb{B}$ denotes the support of $\mathbb{G}_0$. Specifically, $\mathbb{G}_0$ is a Gaussian random element in $\mathbb{B}$ such that $(b_1^*, \ldots, b_K^*) \circ \mathbb{G}_0$ is a centered Gaussian random vector with $\mathbb{C}ov(b_i^*(\mathbb{G}_0), b_j^*(\mathbb{G}_0)) = \mathbb{E}(\tilde{\theta}_{b_i^*}\tilde{\theta}_{b_j^*})$ for all $i, j = 1, \ldots, K$, and $S(\mathbb{G}_0)$ is equal to the closure of $\theta_0'(T(P))$ in $\mathbb{B}$.

Corollary 1.1.1 (Lower Bound for Euclidean Parameters)**.** *Let Assumptions 1.2.1, 1.3.1, 1.3.2, 1.3.3, and assume that the infimum in the display below can be attained. Consider $\theta \in \mathbb{R}^{d_\theta}$ and $\phi \in \mathbb{R}^{d_\phi}$. Then, for any estimator sequence of the form (1.15),*

$$\sup_{I \subset T(P)} \liminf_{n\to\infty} \sup_{h \in I} \mathbb{E}_{P_{n,h}} \left\{ l\left(\sqrt{n}(\hat{\phi}_n - \phi(\theta(P_{n,h}))) \right) \right\} \geqslant \inf_{(v_1,v_2)\in\mathbb{R}^{d_\phi+d_\theta}} \sup_{s \in R(\Sigma_\theta)} \mathbb{E}\left\{ l\left(\phi_0'(\mathbb{G}_0 + s + v_1) - \phi_0'(s) + v_2 \right) \right\},$$

where I is an arbitrary finite subset of the tangent set $T(P)$, $\mathbb{G}_0 \sim N(0, \Sigma_\theta)$ denotes the distributional limit of the efficient (best regular) estimator sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$, and $R(\Sigma_\theta)$ denotes the range of the efficient covariance matrix Σ_θ.

Several comments are in order. First, if the function ϕ is fully differentiable at θ_0, the

lower bound simplifies as follows:

$$\begin{aligned}\inf_{v_1,v_2}\sup_{s}\mathbb{E}\left\{l\left(\phi_0'(\mathbb{G}_0+v_1+s)-\phi_0'(s)+v_2\right)\right\} &\overset{(a)}{=} \inf_{v_1,v_2}\mathbb{E}\left\{l\left(\phi_0'(\mathbb{G}_0)+\phi_0'(v_1)+v_2\right)\right\}\\ &= \inf_{v\in\mathbb{D}}\mathbb{E}\left\{l\left(\phi_0'(\mathbb{G}_0)+v\right)\right\}\\ &\overset{(b)}{=} \mathbb{E}\left\{l\left(\phi_0'(\mathbb{G}_0)\right)\right\},\end{aligned}$$

where (a) follows from the linearity of ϕ_0', and (b) follows from the Anderson's Lemma, since $\phi_0'(\mathbb{G}_0)$ is Gaussian. The expression $\mathbb{E}\left\{l\left(\phi_0'(\mathbb{G}_0)\right)\right\}$ is the well-known risk lower bound for differentiable parameters (e.g., van der Vaart and Wellner, 1996). It implies, in particular, that the "plug-in" estimator $\phi(\hat{\theta}_n)$ is Locally Asymptotically Minimax for any sub-convex loss function. In contrast, the lower bound in Theorem 1.1 suggests that with directionally differentiable functions ϕ, the optimal estimator of the form (1.15) will generally depend on the chosen loss function.

Second, the min-max form of the lower bound is not surprising. The supremum appears by construction, because the theorem deals with the locally maximum risk. In turn, the infimum appears because the lower bound must hold for a large class of competing estimators.

Finally, the lower bound for Euclidean parameters takes a somewhat simpler form. Specifically, note that in Theorem 1.1, the supremum is taken over the support of $\mathbb{G}_0$, which is equal to the closure of the image of the tangent set under the path-wise derivative mapping. If the tangent set is restricted in a complicated way, this set may be hard to characterize. In contrast, the range of the efficient covariance matrix Σ_θ is a relatively simple object. In particular, if Σ_θ is of full rank, $R(\Sigma_\theta)=\mathbb{R}^{d_\theta}$.

Remark 1.1. To study the estimators attaining the lower bound, it will be necessary to work with bounded loss functions, because an application of the Portmanteau lemma is required to establish the distributional convergence of the candidate estimator uniformly over finite neighborhoods of P. To this end, let l be a loss function satisfying Assumption 1.3.3, and l_M be a sequence of bounded, Lipschitz-continuous sub-convex loss functions, converging to l poitwise monotonically from below. For instance, if l is continuous, one can simply

take $l_M = \min\{l, M\}$ for M large enough (Lemma 1.6 in the Appendix provides a general construction). Then, in the notation of Theorem 1.1, the lower bound also holds in the following sense:

$$\lim_{M\to\infty} \sup_{I\subset T(P)} \liminf_{n\to\infty} \sup_{h\in I} \mathbb{E}_{P_{n,h}} \left\{ l_M \left(\sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h))) \right) \right\}$$
$$\geqslant \inf_{(v_1,v_2)\in\mathbb{B}\times\mathbb{D}} \sup_{s\in S(\mathbb{G}_0)} \mathbb{E}\left\{ l\left(\phi_0'(\mathbb{G}_0 + s + v_1) - \phi_0'(s) + v_2 \right) \right\}.$$

1.4.1 Examples Revisited

Example 1 (Continued)**.** Suppose, for simplicity, that $\theta_0 = \mathbb{E}_P(X) \in \mathbb{R}^2$, and the model $\mathbf{P}$ is unrestricted, and focus on the upper bound $\phi(\theta_0) = \min(\theta_{0,1}, \theta_{0,2})$. Then, the sample average $\hat{\theta}_n = n^{-1}\sum_{i=1}^n X_i$ is the best regular estimator, and $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow Z$, where $Z \sim N(0, \Sigma)$ with $\Sigma = \mathbb{V}ar(X)$. Assume that Σ is full rank.

First, consider the binding case when $\theta_{0,1} = \theta_{0,2}$ so that $\phi_0'(h) = \min\{h_1, h_2\}$. The risk lower bound with the quadratic loss $l(x) = x^2$ is given by

$$\inf_{\substack{(v_{11},v_{12})\in\mathbb{R}^2 \\ v_2\in\mathbb{R}}} \sup_{(s_1,s_2)\in\mathbb{R}^2} \mathbb{E}\left\{ (\min(Z_1 + v_{11} + s_1, Z_2 + v_{12} + s_2) - \min(s_1, s_2) + v_2)^2 \right\}$$
$$= \inf_{(v_1,v_2)\in\mathbb{R}^2} \sup_{(s_1,s_2)\in\mathbb{R}^2} \mathbb{E}\left\{ (\min(Z_1 + v_1 + s_1, Z_2 + v_2 + s_2) - \min(s_1, s_2))^2 \right\}.$$

In contrast, when $\theta_{0,1} < \theta_{0,2}$, the derivative is given by $\phi_0'(h) = h_1$, and the risk lower bound simplifies to

$$\inf_{v_1\in\mathbb{R}^2, v_2\in\mathbb{R}} \sup_{(s_1,s_2)\in\mathbb{R}^2} \mathbb{E}\left\{ ((Z_1 + v_{11} + s_1) - (s_1) + v_2)^2 \right\} = \inf_{v\in\mathbb{R}} \mathbb{E}\{(Z_1 - v)^2\} = \mathbb{E}\{Z_1^2\}.$$

The case when $\theta_{0,2} < \theta_{0,1}$ is symmetric. ■

Example 3 (Continued)**.** Suppose again that $N = 2$. Let $\hat{\theta}_n = (\psi_1(\hat{G}_{1:2}), \psi_2(\hat{G}_{2:2}))$ where $\hat{G}_{j:2}$, for $j = 1, 2$ are the empirical CDFs of order statistics of bids. Under suitable assumptions, it can be shown that the model $\mathbf{P}$ is unrestricted. Therefore, $\hat{G}_{1:2}, \hat{G}_{2:2}$ are best regular estimators for $G_{1:2}, G_{2:2}$, and, since ψ_1 and ψ_2 are fully Hadamard differentiable, $\hat{\theta}_n$

is the best regular estimator for θ_0. Moreover, $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a tight centered Gaussian element $\mathbb{G}_0$ in $D([\underline{v}, \overline{v}], [0,1])^2$, which is a vector of Brownian bridges supported on $S(\mathbb{G}_0) = C([\underline{v}, \overline{v}])^2$, where $C([\underline{v}, \overline{v}])$ denotes a set of continuous functions on $[\underline{v}, \overline{v}]$. As in the preceding example, one can verify that the second adjustment term v_2 is not required. Then, for any loss function (e.g., $l(x) = \sup_{v \in [\underline{v}, \overline{v}]} |x(v)|$, or $l(x) = \sum_{j=1}^{d} x(v_j)^2$ for $v_1, \ldots, v_d \in [\underline{v}, \overline{v}]$), the risk lower bound is given by

$$\inf_{w \in D([\underline{v}, \overline{v}], [0,1])^2} \sup_{s \in C([\underline{v}, \overline{v}])^2} \mathbb{E}\left\{ l\left(\phi_0'(\mathbb{G}_0 + w + s) - \phi_0'(s) \right) \right\},$$

where the directional derivative is given in Equations (1.23)–(1.24). ■

1.5 Attaining the Lower Bound

Theorem 1.1 and Remark 1.1 verify that the LAM risk of any estimator of the form (1.15) is bounded from below by:

$$\inf_{(v_1, v_2) \in \mathbb{B} \times \mathbb{D}} \sup_{s \in S(\mathbb{G}_0)} \mathbb{E}\left\{ l_M\left(\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2 \right) \right\}. \tag{1.16}$$

A natural way of obtaining the optimal adjustment terms $(\hat{v}_{1,n}, \hat{v}_{2,n})$ is by minimizing a suitable sample analog of (1.16), as discussed below.

1.5.1 Setup and Assumptions

Denote the population criterion function by

$$Q(v_1, v_2) = \sup_{s \in S(\mathbb{G}_0)} \mathbb{E}\left\{ l_M\left(\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2 \right) \right\}.$$

To construct a sample analog, one has to estimate two unknown components: the distribution of $\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2$ and the support $S(\mathbb{G}_0)$. The law of $\mathbb{G}_0$ can typically be approximated by bootstrap or simulation, so the main complication here is that the directional derivative ϕ_0' is an unknown and potentially non-linear function. Letting $\hat{\mathbb{G}}_n^*$

denote a bootstrap process approximating $\mathbb{G}_0$ and $\hat{\phi}'_n$ denote a suitable estimator for the directional derivative ϕ'_0, the analogy principle suggests approximating the distribution of $\phi'_0(\mathbb{G}_0 + v_1 + s) - \phi'_0(s) + v_2$ by the finite-sample distribution of $\hat{\phi}'_n(\hat{\mathbb{G}}^*_n + v_1 + s) - \hat{\phi}'_n(s) + v_2$ conditional on the data. Next, since $\mathbb{G}_0$ is tight, its support is separable and can be approximated by a sequence of compact sieves. It is not a substantial loss of generality to assume that $\mathbb{G}_0$ is non-degenerate, in which case the support is typically known, but more generally it has to be estimated. Let $(R_n)_{n\geqslant 1}$ denote a sequence of sieves approximating $S(\mathbb{G}_0)$ and $(\hat{R}_n)_{n\geqslant 1}$ denote the corresponding estimators. Then, I choose $(\hat{v}_{1,n}, \hat{v}_{2,n})$ to minimize:

$$\hat{Q}_n(v_1, v_2) = \sup_{s\in\hat{R}_n} \mathbb{E}\left\{ l_M\left(\hat{\phi}'_n(\hat{\mathbb{G}}^*_n + v_1 + s) - \hat{\phi}'_n(s) + v_2\right) \;\middle|\; X_1^n \right\},$$

where the expectation is taken with respect to the distribution of $\hat{\mathbb{G}}^*_n$ conditional on the data. To ensure that $(\hat{v}_{1,n}, \hat{v}_{2,n})$ converge in probability to some minimizers of Q, it is necessary to guarantee that $\hat{Q}_n$ converges to Q uniformly on compact sets. The estimators for the unknown components of Q must be chosen accordingly.

First, I assume that the law of $\mathbb{G}_0$ can be consistently estimated by bootstrap or simulation. Recall that $\mathbb{G}_0$ denotes the distributional limit of the efficient estimator sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$. Let $\hat{\theta}^*_n$ denote the bootstrapped version of $\hat{\theta}_n$, mapping the data X_1^n and bootstrap weights W_1^n, independent of the data, into $\mathbb{B}$. This definition includes nonparametric, Bayesian, block, multiplier and general weighted bootstrap as special cases. Define the set:

$$BL_1(\mathbb{B}) = \left\{ f : \mathbb{B} \to \mathbb{R} : \sup_{b\in\mathbb{B}} |f(b)| \leqslant 1,\ |f(b_1) - f(b_2)| \leqslant ||b_1 - b_2||_{\mathbb{B}} \text{ for } b_1, b_2 \in \mathbb{B} \right\}.$$

Assumption 1.5.1 (Bootstrap Consistency)**.**

(i) $\hat{\theta}^*_n : (X_1^n, W_1^n) \to \mathbb{B}$ *with* W_1^n *independent of* X_1^n *satisfies*

$$\sup_{f\in BL_1(\mathbb{B})} \left| \mathbb{E}(f(\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n))|X_1^n) - \mathbb{E}(f(\mathbb{G}_0)) \right| = o_P(1)$$

under $P_n = \prod_{i=1}^n P$.

(ii) $\sqrt{n}(\hat{\theta}^*_n - \hat{\theta}_n)$ *is asymptotically measurable (jointly in* X_1^n, W_1^n*).*

Condition (i) states that the limiting law of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ can be approximated by the law of $\hat{\mathbb{G}}_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$, conditional on the data.[13] Condition (ii) is a mild measurability assumption that ensures that the bootstrap process converges to $\mathbb{G}_0$ unconditionally.

Second, I assume that the directional derivative can be estimated uniformly well.

Assumption 1.5.2 (Estimating the Directional Derivative)**.** *The estimator $\hat{\phi}'_n : X_1^n \to \mathbb{D}$ of ϕ'_0 satisfies, for any $\delta > 0$,*

$$\sup_{s \in R_n^\delta} \left|\left| \hat{\phi}'_n(s) - \phi'_0(s) \right|\right|_{\mathbb{D}} = o_P(1),$$

where $R_n^\delta = \{b \in \mathbb{B} : d(b, R_n) \leqslant \delta\}$ and $(R_n)_{n \geqslant 1} \subset S(\mathbb{G}_0)$ is an expanding sequence of compact sets.

In view of applying the extremum estimation arguments, the distribution of $\phi'_0(\mathbb{G}_0 + v_1 + s) - \phi'_0(s) + v_2$ must be approximated uniformly in $(v_1, v_2) \in K$ and $s \in R_n$, where K is a fixed compact set and R_n denotes an expanding sequence of compact sets (specified in Assumption 1.5.3). Therefore, the estimator $\hat{\phi}'_n$ must approximate the derivative ϕ'_0 uniformly well. While the above assumption may seem restrictive, natural estimators typically have a stronger property that $\hat{\phi}'_n(b) = \phi'_0(b)$ for all $b \in \mathbb{B}$ with probability approaching one. In practice, such estimators can be based on the analytical expression for ϕ'_0 or obtained by numerical differentiation (see Fang and Santos, 2019; Hong and Li, 2020).

Third, I impose the following assumption on the estimator of the support $S(\mathbb{G}_0)$.

Assumption 1.5.3 (Estimating the Support)**.** *There is an expanding sequence of compact sets $(R_n)_{n \geqslant 1} \subset \mathbb{B}$ such that for any $\varepsilon > 0$ and $s \in S(\mathbb{G}_0)$, there is $s_n \in R_n$ for n large enough such that $||s_n - s|| \leqslant \varepsilon$. The sets R_n are either known or can be estimated with $\hat{R}_n$ satisfying $d_H(\hat{R}_n, R_n) = o_P(1)$ as $n \to \infty$.*

[13] The Bounded Lipchitz distance between two Borel probability measures P and Q is defined as $d_{BL}(P, Q) = \sup_{f \in BL_1} \left| \int f dP - \int f dQ \right|$. It metrizes weak convergence in the sense that a sequence of probability measures P_n converges weakly to a probability measure P if and only if $d_{BL}(P_n, P) = o(1)$ (van der Vaart and Wellner, 1996). Condition (i) can be seen as the sample analog of this requirement, conditional on the data.

Recall that $S(\mathbb{G}_0)$ is equal to the closure of $\theta_0'(T(P))$ in $\mathbb{B}$. Since both θ_0' and $T(P)$ are unknown and the latter may be restricted in a non-trivial way, estimating $S(\mathbb{G}_0)$ is, in general, a complicated task. However, as I show below, Assumption 1.5.3 can be verified in a number of different ways, depending on the application, and does not necessarily require estimating the tangent set $T(P)$ and the path-wise derivative θ_0' directly. See Sections 1.5.2.1 and 1.5.2.2 for further discussion and examples.

Finally, note that the minimization problems with both Q and $\hat{Q}_n$ may have multiple solutions. It is therefore necessary to formulate conditions under which a minimizer of $\hat{Q}_n$ converges in probability to a minimizer of Q.[14] Lemma 1.7 in the Appendix shows that the key requirement for such "point-wise" consistency of the set of minimizers is that $\hat{Q}_n$ converges to Q in probability uniformly over compact sets.

1.5.2 Optimal Estimators

This section contains the second main result of the paper, which develops the optimal estimator of the form (1.15). The result is presented in the general form first and then adapted to a number of special cases.

Theorem 1.2 (Optimal Estimator)**.** *Let Assumptions 1.2.1, 1.3.1 – 1.3.3 and 1.5.1 – 1.5.3 hold and the infimum in the risk lower bound be attained within a compact set $K \subset \mathbb{B} \times \mathbb{D}$. Let $\hat{v}_n = (\hat{v}_{1,n}, \hat{v}_{2,n})$ solve*

$$\inf_{(v_1,v_2)\in K} \sup_{s\in\hat{R}_n} \mathbb{E}\left\{ l_M\left(\hat{\phi}_n'(\hat{\mathbb{G}}_n^* + v_1 + s) - \hat{\phi}_n'(s) + v_2\right) \,\middle|\, X_1^n \right\}, \tag{1.17}$$

where $\hat{\theta}_n$ denotes the efficient (best regular) estimator for θ_0, $\hat{\mathbb{G}}_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ denotes the bootstrap process, and the expectation is taken conditional on the data. Then, the estimator

$$\hat{\phi}_n = \phi\left(\hat{\theta}_n + \frac{\hat{v}_{1,n}}{\sqrt{n}}\right) + \frac{\hat{v}_{2,n}}{\sqrt{n}}$$

[14] More precisely, it suffices to show that $d(\hat{v}_n, \mathcal{V}_0) = o_P(1)$, where $\hat{v}_n = (\hat{v}_{1,n}, \hat{v}_{2,n})$ and $\mathcal{V}_0$ denotes the set of minimizers of Q.

attains the risk lower bound:

$$\lim_{M\to\infty}\sup_{I\subset T(P)}\liminf_{n\to\infty}\sup_{h\in I}\mathbb{E}_{P_{n,h}}\left\{l_M\left(\sqrt{n}(\hat{\phi}_n-\phi(\theta(P_{n,h})))\right)\right\}$$
$$\leqslant \inf_{(v_1,v_2)\in K}\sup_{s\in S(\mathbb{G}_0)}\mathbb{E}\left\{l\left(\phi_0'(\mathbb{G}_0+s+v_1)-\phi_0'(s)+v_2\right)\right\}.$$

Two comments are in order. First, the role and numerical values of the optimal adjustment terms depend on the chosen loss function. In particular, for real-valued parameters $\phi(\theta_0)$, choosing the squared loss function allows to select the adjustment terms that balance the bias-variance trade-off. Second, calculating the optimal adjustment terms amounts to solving the optimization problem in (1.17). This min-max problem may be computationally hard, because the objective function is not convex-concave, and evaluating it at each (v_1, v_2, s) requires bootstrap approximation. However, in many common applications, simple computational heuristics can speed up the optimization, as discussed in Section 1.5.4.

1.5.2.1 Euclidean Parameters

Consider $\theta_0 \in \mathbb{R}^{d_\theta}$ and $\phi(\theta) \in \mathbb{R}^{d_\phi}$. Let Σ_θ denote the variance lower bound for θ and $R(\Sigma_\theta)$ denote its range. According to Corollary 1.1.1 and Remark 1.1, the risk lower bound is given by

$$\inf_{(v_1,v_2)\in\mathbb{R}^{d_\theta+d_\phi}}\sup_{s\in R(\Sigma_\theta)}\mathbb{E}\left\{\, l_M(\phi_0'(\mathbb{G}_0+v_1+s)-\phi_0'(s)+v_2)\,\right\}. \tag{1.18}$$

Let $\hat{\Sigma}_n$ denote a $\sqrt{n}$-consistent estimator of Σ_θ, and σ_j and $\hat{\sigma}_j$ denote the j-th columns of Σ_θ and $\hat{\Sigma}_n$ correspondingly. Define, with $\lambda_n = o(\sqrt{n})$,

$$\begin{aligned} R_n &= \left\{ t=\sum_{j=1}^{d_\theta}\alpha_j\sigma_j\in\mathbb{R}^{d_\theta} \;\middle|\; ||\alpha||\leqslant\lambda_n \right\}, \\ \hat{R}_n &= \left\{ t=\sum_{j=1}^{d_\theta}\alpha_j\hat{\sigma}_j\in\mathbb{R}^{d_\theta} \;\middle|\; ||\alpha||\leqslant\lambda_n \right\}. \end{aligned} \tag{1.19}$$

Then, $\hat{R}_n$ and R_n satisfy Assumption 1.5.3, and the following Corollary holds.

Corollary 1.2.1 (Optimal Estimation of Euclidean Parameters)**.** *Consider* $\theta_0 \in \mathbb{R}^{d_\theta}$, $\phi : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_\phi}$. *Let Assumptions 1.2.1, 1.3.1 - 1.3.3, 1.5.1 (i), and 1.5.2 hold with* $\mathbb{B} = \mathbb{R}^{d_\theta}$ *and* $\mathbb{D} = \mathbb{R}^{d_\phi}$*; define* $\hat{R}_n$ *as in Equation (1.19). Assume that the infimum in (1.18) is attained within a compact set* $K \subseteq \mathbb{R}^{d_\theta + d_\phi}$ *and let* $(\hat{v}_{1,n}, \hat{v}_{2,n})$ *solve*

$$\inf_{(v_1,v_2)\in K} \sup_{s\in\hat{R}_n} \mathbb{E}\left\{ l_M(\hat{\phi}'_n(\mathbb{G}^*_n + v_1 + s) - \hat{\phi}'_n(s) + v_2) \middle| X_1^n \right\}. \tag{1.20}$$

If Σ_θ *is full-rank, the supremum in (1.20) can be taken over* $\mathbb{R}^{d_\theta}$*. Then, the estimator sequence*

$$\hat{\phi}_n \equiv \phi\left(\hat{\theta}_n + \frac{\hat{v}_{1,n}}{\sqrt{n}}\right) + \frac{\hat{v}_{2,n}}{\sqrt{n}}$$

attains the risk lower bound:

$$\lim_{M\to\infty} \sup_{I_f \subset T(P)} \liminf_{n\to\infty} \sup_{h\in I_f} \mathbb{E}_{P_{n,h}} \left\{ l_M\left(\sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h)))\right)\right\}$$
$$\leqslant \inf_{(v_1,v_2)\in\mathbb{R}^{d_\theta+d_\phi}} \sup_{s\in R(\Sigma_\theta)} \mathbb{E}\left\{ l\left(\phi'_0(Z + s + v_1) - \phi'_0(s) + v_2\right)\right\}$$

1.5.2.2 Infinite-Dimensional Parameters

Next, consider estimating the support $S(\mathbb{G}_0)$ according to Assumption 1.5.3 in the settings when $\theta \in \mathbb{B}$ is infinite-dimensional. I will discuss two different approaches.

The first approach is "brute-force" and uses the fact that $S(\mathbb{G}_0)$ equals the closure of $\theta'_0(T(P))$ in $\mathbb{B}$. Let $g_1, g_2, \ldots$ denote a complete sequence in $L_2(P)$, in a sense that for any $f \in L_2(P)$ and any $\varepsilon > 0$, there exist an $m \in \mathbb{N}$, and $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ such that $||f - \sum_{j=1}^m \alpha_j g_j||_{2,P} < \varepsilon$. For example, the space of continuous functions supported on compact sets is dense in $L_2(P)$, and the space of polynomials is dense within that space, by the Stone-Weierstrass theorem. Therefore, $g_1, g_2, \ldots$ can be chosen as properly truncated polynomials. The idea is to use the g_j-s to construct a sequence of compact sieves in the closure of $\theta'_0(T(P))$. To illustrate, suppose that $T(P) = L_2^0(P)$. Let $h_j = g_j - \mathbb{E}_P(g_j)$ denote the projection of g_j onto $L_2^0(P)$, and $\hat{h}_j = g_j - n^{-1}\sum_{i=1}^n g_j(X_i)$ be its sample analog. Let $\hat{\theta}'_n : L_2^0(P) \to \mathbb{B}$ be a suitable estimator for the path-wise derivative map, and define, for

$l_n \in \mathbb{N}$ and $\lambda_n \in \mathbb{R}_+$, the sets

$$\begin{aligned} \hat{R}_n &= \left\{ \sum_{j=1}^{l_n} \alpha_j \hat{\theta}'_n(\hat{h}_j) \;\middle|\; ||\alpha|| \leqslant \lambda_n \right\}, \\ R_n &= \left\{ \sum_{j=1}^{l_n} \alpha_j \theta'_0(h_j) \;\middle|\; ||\alpha|| \leqslant \lambda_n \right\}. \end{aligned} \tag{1.21}$$

The following Lemma provides primitive conditions under which $\hat{R}_n$ and R_n defined above satisfy Assumption 1.5.3.

Lemma 1.1 (Estimating the Support via Projections)**.** *Assume that:*

1. $\left|\left|\hat{\theta}'_n(1) - \theta'_0(1)\right|\right|_{\mathbb{B}} = o_P(1)$ *and* $\lambda_n \cdot \max_{j \leqslant l_n} \left|\left|\hat{\theta}'_n(g_j) - \theta'_0(g_j)\right|\right|_{\mathbb{B}} = o_P(1)$
2. $\lambda_n \cdot \sqrt{\frac{l_n}{n}} \cdot \max_{j \leqslant l_n} ||g_j||_{2,P} = o(1)$

Then $\hat{R}_n$ and R_n defined in (1.21) satisfy Assumption 1.5.3.

Assumption 1 is a point-wise and uniform consistency requirement on $\hat{\theta}_n$, which can be verified via a suitable maximal inequality or with the sample splitting technique. Assumption 2 is a rate condition, which relates the number and "size" of elements in the construction of R_n with n. Similar primitive conditions can be formulated in settings where the tangent set $T(P)$ is restricted.

The second approach is similar in spirit to the Euclidean case, and is based on characterizing the support of a Gaussian process $\mathbb{G}_0$ via its covariance kernel. The main idea is illustrated below in the example where $\mathbb{G}_0$ is a Gaussian process with $S(\mathbb{G}_0) = C_b([0,1])$ endowed with the sup-norm. The technical details are deferred to Remark 1.2. Let $K : [0,1] \times [0,1] \to \mathbb{R}$ defined by $K(s,t) = \mathbb{E}(\mathbb{G}_0(t)\mathbb{G}_0(s))$ denote the covariance kernel of $\mathbb{G}_0$,

and $\hat{K}_n$ denote a suitable estimator. Denote:

$$R_n = \left\{ f(s) = \sum_{j=1}^{l_n} \alpha_j K(t_j, s) \;\middle|\; 0 \leqslant t_1 < \cdots < t_{l_n} \leqslant 1;\ ||\alpha|| \leqslant \lambda_n \right\},$$
$$\hat{R}_n = \left\{ f(s) = \sum_{j=1}^{l_n} \alpha_j \hat{K}(t_j, s) \;\middle|\; 0 \leqslant t_1 < \cdots < t_{l_n} \leqslant 1;\ ||\alpha|| \leqslant \lambda_n \right\}. \tag{1.22}$$

The following Lemma provides a primitive condition under which $\hat{R}_n$ and R_n satisfy Assumption 1.5.3.

Lemma 1.2 (Estimating the Support via Covariance Kernel)**.** *Let R_n and $\hat{R}_n$ be defined in (1.22) with $l_n \in \mathbb{N}$, and $\lambda_n \in \mathbb{R}_+$. Suppose that $\hat{K}_n : [0,1] \times [0,1] \to \mathbb{R}$ satisfies*

$$\lambda_n \max_{j \leqslant l_n} ||\hat{K}_n(t_j, \cdot) - K(t_j, \cdot)||_\infty = o_P(1).$$

Then, R_n and $\hat{R}_n$ satisfy Assumption 1.5.3.

Remark 1.2 (Support of a Gaussian Measure and Cameron-Martin Space)**.** The exposition below follows Bogachev (1998). Since $\mathbb{G}_0$ is tight, it concentrates on the separable subspace of $\mathbb{B}$, which I denote $\mathbb{B}_0$, and induces a centered Radon Gaussian measure γ on $(\mathbb{B}_0, \mathcal{B}(\mathbb{B}_0))$ (see Theorem 7.1.7. in Bogachev, 2007). The support of $\mathbb{G}_0$ is equal to the closure of $H(\gamma)$ in $\mathbb{B}_0$, where $H(\gamma)$ denotes the Cameron-Martin space of γ, constructed as follows (Theorem 3.6.1. in Bogachev, 1998). Each element of the continuous dual $\mathbb{B}_0^*$ is a Normal random variable defined on $(\mathbb{B}_0, \mathcal{B}(\mathbb{B}_0), \gamma)$. This allows to view $\mathbb{B}_0^*$ as a subset of $L_2(\gamma)$. Let $\mathbb{B}_\gamma^*$ denote the $L_2(\gamma)$-closure of $\mathbb{B}_0^*$. For each $h \in \mathbb{B}_0$, let $L_h : \mathbb{B}_\gamma^* \to \mathbb{R}$ denote the evaluation map $L_h(b^*) = b^*(h)$. The Cameron-Martin space of γ is defined as $H(\gamma) = \{h \in \mathbb{B}_0 : L_h \text{ is continuous w.r.t } ||\cdot||_{2,\gamma}\}$. Next, for each $b^* \in \mathbb{B}_\gamma^*$, let $K(b^*, \cdot) : \mathbb{B}_\gamma^* \to \mathbb{R}$ be defined by

$$K(b^*, c^*) = \int_{\mathbb{B}} b^*(x) c^*(x) d\gamma(x) = \mathbb{E}(b^*(\mathbb{G}_0) c^*(\mathbb{G}_0)).$$

By Theorem 3.2.3 in Bogachev (1998), for each $b^* \in \mathbb{B}_\gamma^*$, there is $h_{b^*} \in H(\gamma)$ such that $K(b^*, c^*) = c^*(h_{b^*})$ for all $c^* \in \mathbb{B}_\gamma^*$. In this sense, every element of $\mathbb{B}_\gamma^*$ can be associated with

a unique element of $H(\gamma)$. Therefore, the set $H(\gamma)$ can be mapped out by choosing different b^* and finding the associated h_b^*.

For example, let $\mathbb{B}_0 = C_b([0,1])$ be a set of continuous bounded functions on $[0,1]$, and $\mathbb{G}_0$ denote a Gaussian process with covariance kernel $K(s,t) \equiv \mathbb{E}(\mathbb{G}_0(s)\mathbb{G}_0(t))$. Recall that the continuous dual $\mathbb{B}_0^*$ is the set of all finite Borel measures on $[0,1]$ so that $b^*(x) = \int x(t)d\mu_{b^*}(t)$. With the help of Fubini's theorem, one can verify that $h_{b^*}(s) = \int K(s,t)d\mu_{b^*}(t)$. Further, the set of finitely-supported Borel measures $\{\sum_{j=1}^{J} \alpha_j \delta_{t_j} : \alpha_j \in \mathbb{R}, t_j \in [0,1], J \in \mathbb{N}\}$, where δ_t denotes the Dirac measure with mass at t, is weak-star dense in $\mathbb{B}_0^*$ meaning that any such $h_{b^*}(s)$ can be approximated by a sequence of the form $\sum_j \alpha_j K(s,t_j)$ point-wise in s, and therefore uniformly since $s \in [0,1]$. This motivates the definition of R_n in (1.22).

1.5.3 Examples Revisited

Example 1. Focus on the upper bound $\phi(\theta_0) = \min_{j \leqslant d}(\theta_{0,j})$ with $\theta_0 \in \mathbb{R}^d$. Here, estimating the directional derivative (see Equation 1.7) amounts to selecting θ_j that are sufficiently close to each other, which is essentially an inequality selection problem. One way to proceed is to test a set of hypotheses $H_0 : \theta_{0,j} \leqslant \theta_{0,i}$ for all i, j (following e.g. Romano et al. (2014)), collect all j-s for which the null is not rejected into the set $\hat{B}_n$, and set

$$\hat{\phi}'_n(h) = \min_{j \in \hat{B}_n}(h_j)$$

Then, if the test size approaches zero as n approaches infinity, $\hat{\phi}'_n(h) = \phi'_0(h)$ for all $h \in \mathbb{R}^d$, with probability approaching one, so that the resulting estimator satisfies Assumption 1.5.2. ■

Example 3. Suppose again that $N = 2$, and let $\hat{\theta}_n = (\psi_1(\hat{G}_{1:2}), \psi_2(\hat{G}_{2:2}))$ where $\hat{G}_{j:2}$, for $j = 1, 2$ are the empirical CDFs of order statistics of bids. The form of the directional derivative in Equation (1.24) suggests a natural sample counterpart. For a positive sequence

$\kappa_n \downarrow 0$, define the sets

$$\begin{aligned}\hat{S}_{1,n} &= \{v \in [\underline{v}, \overline{v}] : \psi_1(\hat{G}_{1:2}(v)) < \psi_2(\hat{G}_{2:2}(v)) - \kappa_n\},\\ \hat{S}_{2,n} &= \{v \in [\underline{v}, \overline{v}] : \psi_2(\hat{G}_{2:2}(v)) < \psi_1(\hat{G}_{1:2}(v)) - \kappa_n\},\\ \hat{S}_{0,n} &= \{v \in [\underline{v}, \overline{v}] : |\psi_1(\hat{G}_{1:2}(v)) - \psi_2(\hat{G}_{2:2}(v))| \leqslant \kappa_n\},\end{aligned} \tag{1.23}$$

and set, for any $h \in D([\underline{v}, \overline{v}], [0,1])^2$,

$$\hat{\phi}'_n(h)(v) = h_1(v)\mathbf{1}(v \in \hat{S}_{1,n}) + h_2(v)\mathbf{1}(v \in \hat{S}_{2,n}) + \min(h_1(v), h_2(v))\mathbf{1}(v \in \hat{S}_{0,n}). \tag{1.24}$$

Then, if $\kappa_n\sqrt{n} \to \infty$, one can show that the resulting estimator satisfies Assumption 1.5.2 even with R_n^δ replaced by $D([\underline{v}, \overline{v}], [0,1])^2$. ■

1.5.4 Computation

In some special cases, computation of the adjustment terms can be substantially simplified by splitting the optimization problem into several independent sub-problems or using approximate closed-form solutions. More generally, I discuss computational heuristics that can be applied to speed-up the optimization.

The main factor that slows down the optimization problem in (1.20) is that the objective function is costly to evaluate. The approach discussed below aims to reduce the number of evaluations. I focus on the finite-dimensional parameters for simplicity, but similar ideas can be applied in infinite-dimensional settings as well, after selecting suitable sieves. The lower bound from Corollary 1.1.1 can be equivalently written as

$$\inf_{(v_1,v_2)\in\mathbb{R}^{d_\theta+d_\phi}} \sup_{s\in B} \sup_{\lambda\geqslant 0} \mathbb{E}\left\{l\left(\phi'_0(Z+\lambda s+v_1) - \lambda\phi'_0(s) + v_2\right)\right\}, \tag{1.25}$$

where B denotes the unit ball in $\mathbb{R}^{d_\theta}$. For a fixed v_1, v_2 and s, consider a function

$$g(\lambda) = \mathbb{E}\left\{l\left(\phi'_0(Z+\lambda s+v_1) - \lambda\phi'_0(s) + v_2\right)\right\}$$

that traces the value of the objective function along the ray passing through s. A useful property that appears to hold in practice but turns out to be hard to prove theoretically is

that $g(\lambda)$ is maximized at zero or infinity. Therefore, for each (v_1, v_2), the supremum can be calculated by selecting a set of directions (i.e., values of s) on the unit ball and evaluating the function $g(\lambda)$ at zero and some large value of the argument in each direction. Since the directional derivative is typically a partially linear function with a small number of different slopes, this approach allows to reduce the number of evaluations of the objective function dramatically.

In special cases, such as $\phi_0'(h) = \max_{j \leqslant d}(h_j)$ with the squared loss function, following the above line of thought allows to formulate an approximate closed-form solution. In such cases, $v_2 = 0$ without loss of generality (for any loss function). Imposing an additional assumption that $v_1 = (v, \ldots, v)^T \in \mathbb{R}^d$ and elaborating on the arguments above suggests the folowing solution

$$v^* = \frac{1}{2} \max_{I \subset \{1,\ldots,d\}} \left(\frac{\mathbb{E}((\max_{j \in I} Z_i)^2) - \mathbb{E}(Z_{i^*}^2)}{\mathbb{E}(\max_{j \in I} Z_j)} \right), \tag{1.26}$$

where $i^* = \operatorname{argmax}_{i \leqslant d} \mathbb{E}(Z_i^2)$ and the maximum over empty set is set to be equal to zero, which guarantees $v^* \geqslant 0$. Similarly, with $\phi_0'(\theta) = \min_{j \leqslant d}(\theta_j)$ and the squared loss, the solution is given by

$$v^* = \frac{1}{2} \min_{I \subset \{1,\ldots,d\}} \left(\frac{\mathbb{E}((\min_{j \in I} Z_i)^2) - \mathbb{E}(Z_{i^*}^2)}{\mathbb{E}(\min_{j \in I} Z_j)} \right). \tag{1.27}$$

Extensive simulations suggest that these closed-form adjustment terms actually attain the global minimum in (1.25), although the corresponding formal result is hard to establish. Recalling that $Z \sim N(0, \Sigma)$ with Σ consistently estimated by $\hat{\Sigma}_n$ suggests the following procedure: (i) draw $Z_1^*, \ldots, Z_B^* \sim N(0, \hat{\Sigma}_n)$ for some large B and (ii) replace expectations with sample averages in the expressions above.

The above formulas can be applied in other settings as well. Consider Example 1 with $\theta = (\theta_1, \theta_2) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ that do not have any common components and $\phi_0' : \mathbb{R}^{d_1+d_2} \to \mathbb{R}^2$ given by $\phi_0'(h) = (\min_{j \leqslant d_1}(h_{1,j}), \max_{k \leqslant d_2}(h_{2,k}))^T$. Then, with the quadratic loss function $l : \mathbb{R}^2 \to \mathbb{R}_+$ defined as $l(x_1, x_2) = x_1^2 + x_2^2$, the optimization problem can be separated into

two independent subproblems:

$$\inf_{(v_1,v_2)\in\mathbb{R}^{d_1+d_2+2}} \sup_{s\in\mathbb{R}^{d_1+d_2}} \mathbb{E}\left\{l\left(\phi_0'(Z+s+v_1)-\phi_0'(s)+v_2\right)\right\}$$

$$= \inf_{(v_{11},v_{12})\in\mathbb{R}^{d_1+1}} \sup_{s_1\in\mathbb{R}^{d_1}} \mathbb{E}\left\{\left(\min(Z_1+s_1+v_{11})-\min(s_1)+v_{12}\right)^2\right\}$$

$$+ \inf_{(v_{21},v_{22})\in\mathbb{R}^{d_2+1}} \sup_{s_2\in\mathbb{R}^{d_1}} \mathbb{E}\left\{\left(\max(Z_2+s_2+v_{21})-\max(s_2)+v_{22}\right)^2\right\}.$$

Then, $v_{12}^* = v_{22}^* = 0$ and the approximate solutions (v_{11}^*, v_{21}^*) to each of the problems are given by equations (1.27) and (1.26) correspondingly. Similar arguments can be applied in the setting of Example 3 if the loss function $l : D([\underline{v}, \overline{v}]) \to \mathbb{R}_+$ is given by $l(x) = \sum_{i=1}^d x(v_i)^2$ for some fixed $v_1, \ldots, v_d \in [\underline{v}, \overline{v}]$.

1.6 Simulation Study

I illustrate the finite-sample performance of the proposed estimator by comparing it with the simple "plug-in" estimator and the existing bias correction approaches. For simplicity, I focus on the upper bound from Example 1: $\phi(\theta) = \min_{j\leqslant d}(\theta_j)$ with $\theta \in \mathbb{R}^d$. The results for the lower bound and for both bounds together are similar.

I start by discussing the existing bias-correction approaches. The first approach, considered in Kreider and Pepper (2007), is to use bootstrap bias correction (Tibshirani and Efron, 1993; Horowitz, 2001). It is implemented as follows: (i) Draw B bootstrap samples $\{X_1^*, \ldots, X_n^*\}$, and calculate $\bar{X}_b^* = \frac{1}{n}\sum_{i=1}^n X_i^*$; (ii) Estimate the bias by $\hat{b}_n^* = \frac{1}{B}\sum_{b=1}^B \phi(\bar{X}_b^*) - \phi(\hat{\theta}_n)$, and compute the adjusted estimator

$$\hat{\phi}_n^{\text{Bootstrap}} \equiv \phi(\hat{\theta}_n) - \hat{b}_n^* = 2\phi(\hat{\theta}_n) - \frac{1}{B}\sum_{b=1}^B \phi(\bar{X}_b^*).$$

Kreider and Pepper (2007) found that this method performs well in practice, even though it is not fully theoretically justified.[15] Studying the asymptotic properties of such estimator is

[15]The standard arguments for consistency of the procedure rely on the differentiability of the function ϕ,

beyond the scope of this paper.

The second approach is due to Chernozhukov et al. (2013). The authors propose a half-median unbiased estimator which lies above the true value with probability at least one half asymptotically.[16] The estimator takes the form

$$\hat{\phi}_n^{\text{CLR}} \equiv \phi(\hat{\theta}_n + \hat{c}_n),$$

where $\hat{c}_n$ is the adjustment term calculated in two steps. The first step performs inequality selection, picking the components of θ_0 that are sufficiently close to each other, and the second step focuses on the selected components to choose the appropriate adjustment term. Although the form of $\hat{\phi}_n^{CLR}$ is very similar to the estimator proposed in this paper, the two approaches are very different. The adjustment term $\hat{c}_n$ is chosen to reduce the bias of the "plug-in" estimator, and may lead to large LAM risk, while the adjustment terms proposed in this paper minimize the risk and do not target the bias directly.

Next, consider the implementation of the proposed estimator. Let $Z \sim N(0, \Sigma)$, denote the weak limit of the efficient estimator sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$. To approximate the law of Z in accord with Assumption 1.5.1, one may pick a consistent estimator $\hat{\Sigma}_n$ for Σ and chose Z_n^* to be a random vector distributed as $N(0, \hat{\Sigma}_n)$, conditional on the data. To construct a suitable estimator for the directional derivative, one may follow the procedure described in Section 1.5.3 and obtain $\hat{\phi}'_n(h) = \min_{j \in \hat{B}_n}(h_j)$. Then, calculate the adjustment term $\hat{v}_{1,n}$ by minimizing

$$\inf_{v_1 \in \mathbb{R}^d} \sup_{c \in \mathbb{R}^d} \mathbb{E}\left((\hat{\phi}'_n(Z_n^* + v_1 + c) - \hat{\phi}'_n(c))^2 \middle| X_1^n \right)$$

and set

$$\hat{\phi}^{\text{LAM}} \equiv \phi\left(\hat{\theta}_n + \frac{\hat{v}_{1,n}}{\sqrt{n}}\right).$$

---

which, in the present setting, may fail. See Tibshirani and Efron (1993).

[16] This criterion is considered beacuse the results of Hirano and Porter (2012) suggests that median-unbiased estimators do not exist.

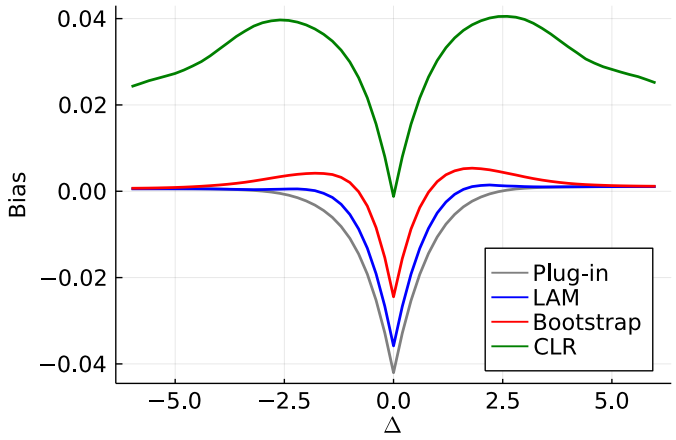
In this special case the second adjustment term is not required and the optimization problem is simplified. Moreover, the squared loss function allows to choose $\hat{v}_{1,n}$ to balance the bias-variance trade-off, and compute approximate closed-form solutions as discussed in Section 1.5.4.

The simulation setup is as follows. The data $X_1, \ldots, X_n$ are i.i.d. from $N(\theta_0, \Sigma)$ in $\mathbb{R}^3$, so that $\theta_0 = \mathbb{E}_P(X)$. I consider an ordinary covariance matrix Σ with different variances and non-zero correlations, and set $\theta(\Delta) = (0, \Delta/\sqrt{n}, 2\Delta/\sqrt{n})^T$ so that Δ plays the role of the local parameter. That is, Δ equal to zero corresponds to the point $\theta_0 = (0,0,0)^T$, where the full differentiability of ϕ fails, and varying Δ allows to "walk across" the local neighborhood of this point.[17] For each value of the local parameter Δ on a grid chosen to scale, I perform $M = 5000$ simulations, with $B = 2000$ bootstrap draws and sample size $n = 300$. For every draw, indexed by m, I generate a random sample $X_1^m, \ldots, X_n^m$ from $N(\theta_0, \Sigma)$, and calculate $\hat{\phi}_m^{\text{Plug-in}} = \phi(\bar{X}_m)$, and $\hat{\phi}_m^{\text{Bootstrap}}$, $\hat{\phi}_m^{\text{CLR}}$ and $\hat{\phi}_m^{\text{LAM}}$ according to the formulas above. Then, I compute the average bias, $\frac{1}{M}\sum_{m=1}^{M}(\hat{\phi}_m - \phi(\theta(\Delta)))$, and risk, $\frac{1}{M}\sum_{m=1}^{M}(\hat{\phi}_m - \phi(\theta(\Delta)))^2$, for each of the four estimators and plot the results as a function of Δ.

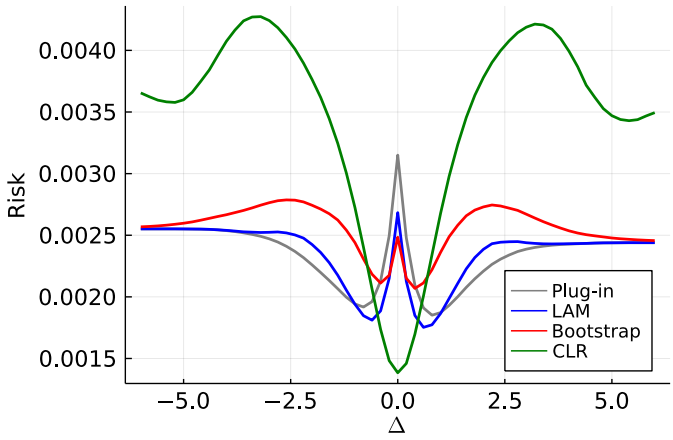
The results presented in Figure 1.2 require several comments. First, Panel (a) suggests that the LAM estimator does not reduce the bias as much as the other methods. This is not surprising, since the LAM estimator was constructed targeting the mean-squared error (i.e., variance plus bias squared), rather than the bias directly. Larger reduction in bias can be achieved by using a different loss function, such as $l(x) = |x|^\alpha$ for $0 < \alpha < 2$. Second, Panel (b) suggests that the LAM estimator has the lowest worst-case risk, which is consistent with the asymptotic results of Theorems 1.1 and 1.2. Note that while the risk of the plug-in estimator is maximized at zero (i.e., at the point of non-differentiability) the maximum risks of the bias-corrected estimators are attained away from zero. Moreover, the LAM estimator outperforms the bias-correction methods in terms of risk everywhere

[17] There are many other curves that pass through $\theta_0 = (0,0,0)$, and this particular choice is made only for illustrative purposes. The last coordinate of θ_0 is multiplied by two only for aesthetic reasons, to ensure that the graphs are symmetric and properly scaled.

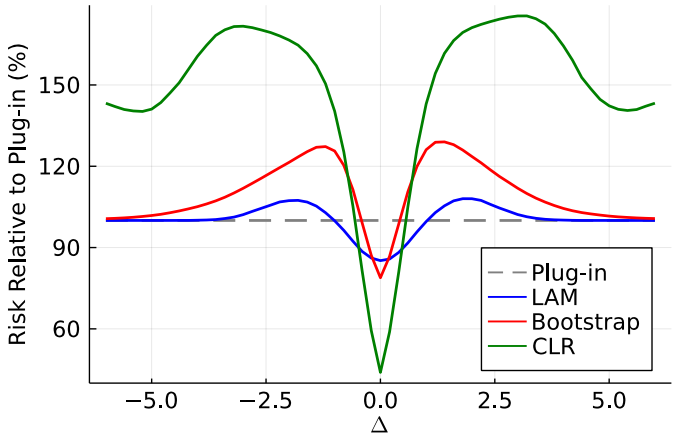
Figure 1.2: Finite-Sample Bias, Risk, and Relative Risk.



(a)



(b)



(c)

Notes: The horizontal axis corresponds to the local parameter Δ. Panels (a) and (b) are in absolute terms. Panel (c) shows the efficiency gains (or losses) of the estimators relative to the "Plug-in" estimator.

except for a small neighborhood of zero. Finally, Panel (c) shows relative risks in percentage terms, suggesting that the bias-corrected estimators may have a substantially larger risk than the Plug-in, depending on the value of Δ, while the LAM estimator does not. Since Δ is unknown and cannot be consistently estimated, the LAM estimator can be interpreted as cautious.

Extensive additional simulations suggest that the amount of bias and risk reduction of the LAM estimator (relative to Plug-in) increase in the dimension of θ, and decrease in the correlation between the components of $\hat{\theta}_n$.

1.7 English Auctions with IPV

In this section, I revisit the model of English auctions with independent private values from Haile and Tamer (2003). I apply the developed theory to construct efficient estimators for the bounds on the distribution of valuations and the implied bounds optimal reserve price, and compare the results with Haile and Tamer (2003). Using empirically calibrated simulations, I find that the proposed estimator, on average, yields substantially sharper bounds.

1.7.1 Model and Identification

Consider a symmetric English auction. Suppose that there are N bidders, and each bidder j draws his valuation $V_j \in [\underline{v}, \overline{v}]$, independently of the others, from a distribution with a cumulative distribution function denoted by F. Let B_j denote the final bid of player j and $B_{j:N}$ denote the j-th lowest final bid in a given auction. Assume that the reserve price is below $\underline{v}$, and let $\Delta > 0$ denote the minimal bid increment.

1.7.1.1 CDF of Valuations

The main primitive parameter of interest in this setting is the marginal distribution of valuations F. The knowledge of this distribution allows to forecast the expected revenue and bidders surplus and study the effects of a counterfactual change in the auction design, such as setting a different reserve price. To relate this distribution with the observed distribution of bids, one has to make assumptions on the bidding behavior. Haile and Tamer (2003) assume that each player: (i) does not bid above his valuation and (ii) does not let the others win at a price he is willing to pay. Assumption (i) states that $B_j \leqslant V_j$ for each $j \leqslant N$, implying that the order statistics satisfy $B_{j:N} \leqslant V_{j:N}$ for each $j \leqslant N$, and

$$F_{j:N}(v) \leqslant G_{j:N}(v),$$

where $F_{j:N}$ and $G_{j:N}$ denote the distributions of the j-th order statistics of valuations and bids correspondingly. Assumption (ii) implies that $V_{N-1:N} \leqslant B_{N:N} + \Delta$, and, therefore,

$$F_{N-1:N}(v) \geqslant G_{N:N}(v - \Delta).$$

It is well-known that the distribution of any order statistic of a collection of i.i.d. random variables uniquely determines the parent distribution: for each $j \leqslant N$, there is a strictly increasing and differentiable function $\psi_j : [0,1] \to [0,1]$ such that $F(v) = \psi_j(F_{j:N}(v))$[18]. Applying ψ_j to both sides of the two previous displays for every $j \leqslant N$ and intersecting the results, Haile and Tamer (2003) obtain the following point-wise bounds:

$$\psi_{N-1}(G_{N:N}(v - \Delta)) \leqslant F(v) \leqslant \min_{j \leqslant N} \psi_j(G_{j:N}(v)). \tag{1.28}$$

While these bounds are not sharp (Chesher and Rosen, 2017), they can be sufficiently informative.

[18]Specifically, $\psi_j(t)$ is defined implicitly through $t = n!/((n-j)!(i-j)!) \int_0^{\psi_i} s^{j-1}(1-s)^{n-j} ds$; see e.g. Arnold et al. (2008).

1.7.1.2 Optimal Reserve Price

One of the main policy variables for the seller is the reserve price. Haile and Tamer (2003) show that, under suitable assumptions on the distribution of valuations and bidding strategies in counterfactual auctions, informative bounds on the optimal reserve price can be obtained directly from the bounds on the distribution of valuations derived above. Specifically, assume that F is strictly increasing and continuously differentiable, and such that the function $\pi(p; F)$ defined below is strictly pseudo-concave. Then, in any feasible auction mechanism that is revenue equivalent to the second-price sealed-bid auction in the sense of Myerson (1981), the optimal reserve price maximizes

$$\pi(p; F) = (p - v_0)(1 - F(p)),$$

where v_0 denotes the value of the unsold good to the seller. Denoting the bounds on the CDF by $F_L(v) \leqslant F(v) \leqslant F_U(v)$, it follows that $\pi(p; F_U) \leqslant \pi(p; F) \leqslant \pi(p; F_L)$ for all p. As illustrated in Figure 1.3, this implies the following bounds $[p_L, p_U]$ on the optimal reserve price:

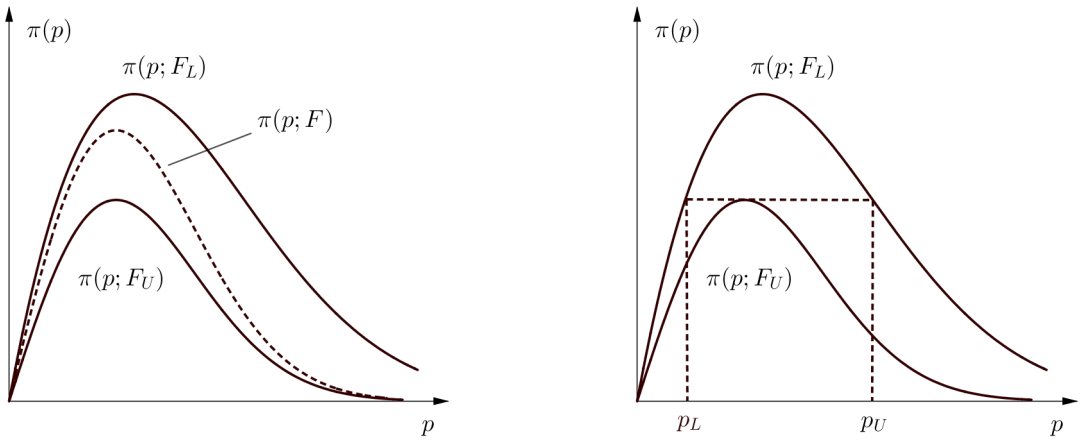
$$\begin{aligned} p_L &= \inf \left\{ p \in [\underline{v}, \overline{v}] : \pi(p; F_L) \geqslant \max_{p' \in [\underline{v}, \overline{v}]} \pi(p'; F_U) \right\}, \\ p_U &= \sup \left\{ p \in [\underline{v}, \overline{v}] : \pi(p; F_L) \geqslant \max_{p' \in [\underline{v}, \overline{v}]} \pi(p'; F_U) \right\}. \end{aligned}$$

Note that, even if the bounds on the CDF of valuations and expected profit are relatively tight, the implied bounds on the optimal reserve price may still be fairly wide.

1.7.2 Estimation

It is assumed that the researcher observes an i.i.d. sample of auction data which includes bids $\{B_i\}_{i=1}^n$ where $B_i = (B_{1,i}, \ldots, B_{N,i})$. Such data can be used to estimate the empirical CDFs of order statistics of bids.[19] Consider estimating the upper bound on the distribution

[19] The analysis can be performed conditional on auction characteristics and the number of participants. To apply the results of this paper, the auction characteristics must be discrete (or discretized) to ensure that the conditional CDF-s of the bids can be regularly estimated (see Section 1.3.2.2). Note that, since the IPV assumption is imposed conditional on the auction characteristics, focusing on discrete characteristics may be restrictive.



(a) Bounds on the true profit function

(b) Implied bounds on the maximizer

Figure 1.3: Identification of the Optimal Reserve Price.

of valuations from Equation (1.28). For a fixed $v \in [\underline{v}, \overline{v}]$, the upper bound takes the form $\phi(\theta(v)) = \min_{j \leqslant d}(\theta_d(v))$, where $\theta(v)$ is a vector of smooth transformations of the CDFs of bids evaluated at v. Haile and Tamer (2003) propose to approximate the minimum by a sequence of smooth functions, chosen to reduce the finite-sample bias. Specifically, they consider the function

$$\tilde{\phi}(\theta; \rho) = \sum_{j=1}^{d} \theta_j \frac{\exp(\rho \cdot \theta_j)}{\sum_{k=1}^{d} \exp(\rho \cdot \theta_k)},$$

where ρ is the smoothness parameter. This function satisfies $\tilde{\phi}(\theta; \rho) > \min_{j \leqslant d}(\theta_j)$ for any $\rho \in \mathbb{R}$, and $\lim_{\rho \to -\infty} \mu(\theta; \rho) = \min_{j \leqslant d}(\theta_j)$. Letting $\hat{\theta}_n$ denote an estimator for θ_0 and $\rho_n \to -\infty$ denote an appropriate sequence of smoothing parameters,[20] they set

$$\hat{\phi}_n^{HT} \equiv \tilde{\phi}(\hat{\theta}_n; \rho_n) = \sum_{j=1}^{J} \hat{\theta}_j \frac{\exp(\rho_n \cdot \hat{\theta}_{j,n})}{\sum_{k=1}^{J} \exp(\rho_n \cdot \hat{\theta}_{k,n})}.$$

Such estimator has the same asymptotic properties as $\hat{\phi}_n^{\text{Plug-in}} = \min_{j \leqslant d}(\hat{\theta}_{j,n})$, with the advantage of providing bias-correction in finite-samples.[21]

[20] To ensure a suitable amount of bias-correction, the sequence should not diverge too fast. On the other hand, it cannot diverge too slow, or the bias will become infinite. Haile and Tamer (2003) derive the asymptotic properties of their estimator with ρ_n diverging faster than $\log \sqrt{n}$.

[21] From the asymptotic efficiency perspective, the two estimators are equivalent.

While the above estimator is computationally simple and provides sufficient bias-correction, it may be inefficient: Attempting to reduce the bias by choosing ρ_n close to zero may disproportionally increase the variance of the resulting estimator. Additionally, this estimator does not account for the fact that $\theta(v)$ is estimated with different precision at different points of the support (unless one somehow selects a different smoothing parameter for each $v \in [\underline{v}, \overline{v}]$). In turn, with a suitable choice of the loss function, the proposed estimator can optimally balance the bias-variance trade-off and automatically adapt to the precision of the estimates of $\theta(v)$. It can also be implemented in computationally simple way and computed within several seconds, as discussed below.

Construction of the proposed estimator in this setting has been discussed in Example 3 throughout the paper. The parameter of interest is a pair of CDF-type functions, $\phi(\theta_0) \in D([\underline{v}, \overline{v}], [0,1])^2$, representing the bounds on F in Equation (1.28). To focus on the bias-variance trade-off in estimation and simplify the computation of the adjustment terms, I consider the squared loss function that focuses on a finite grid of points $v_1, \ldots, v_K \in [\underline{v}, \overline{v}]$. Specifically, the loss function $l : D([\underline{v}, \overline{v}], [0,1])^2 \to \mathbb{R}_+$ is given by $l(x_1, x_2) = \sum_{k=1}^{K} (x_1(v_k)^2 + x_2(v_k)^2)$. Then, as discussed in Section 1.5.4, the optimization problem can be split into several simple subproblems that have approximate closed-form solutions.

1.7.3 Results

I compare the performance of the two estimation methods on simulated data. To mimic the empirical results of Haile and Tamer (2003), the true distribution or valuations is taken to be Log-Normal with parameters $\mu = 4$ and $\sigma = 0.5$, the minimal bid increment is $\Delta = 5$, and jump bids (substantially exceeding the bid increment) are allowed. The bidding process is designed to satisfy Assumptions (i) and (ii) above, and may substantially differ from the standard button auction model. Only the final bid of each participant is recorded.

Figure 1.4 presents the results. First, since the lower bound equals to $\psi_{N-1}(G_{N:N}(v-\Delta))$, no smoothing or adjustment is required and the two estimation methods yield the same

Table 1.1: Estimated Bounds on the Optimal Reserve Price

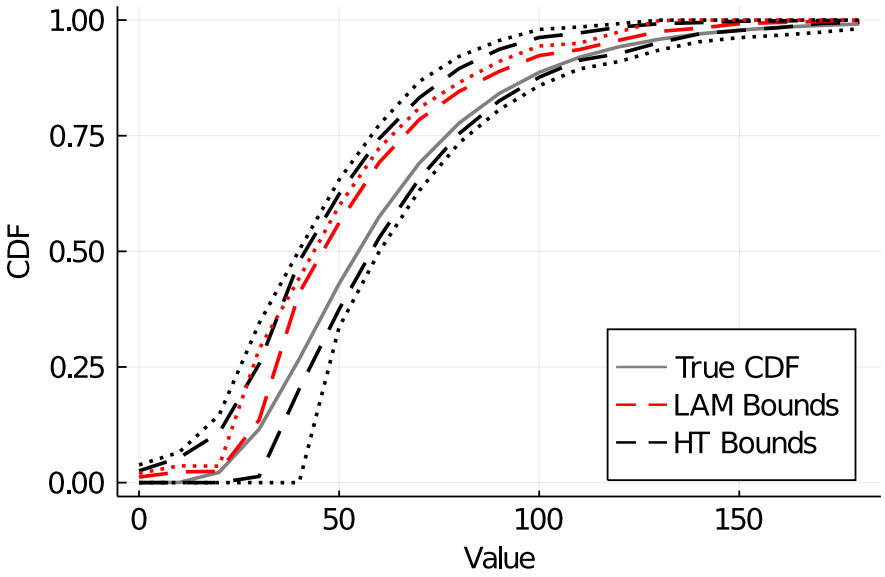
Parameters	$\mu = 4, \sigma = 0.5$	$\mu = 3, \sigma = 1$	$\mu = 5, \sigma = 0.25$
True p^*	42.1	27.2	112.6
$F(p^*)$	0.3	0.62	0.13
Mean LAM bounds	$[34.0, 59.6]$	$[14.8, 75.5]$	$[97.3, 139.3]$
Mean HT bounds	$[27.5, 68.9]$	$[8.3, 84.6]$	$[91.3, 141.4]$
LAM / HT width	61.5%	79.5%	83.3%

Note: Valuations are drawn from the Log-Normal distribution with parameters μ and σ. The number of bidders is $N = 6$, sample size is $n = 200$.

results. The estimated lower bound is fairly tight throughout the support since the minimal bid increment is relatively small and jump bidding is not too common. Second, the LAM estimates for the upper bound are, on average, substantially tighter than the HT estimates. In particular, the 95th quantile for the LAM estimate (red dotted line) is consistently below the average HT estimate (black dashed line) across simulations. At the same time, there is some downward bias in the LAM estimates around the lower part of the support. This issue, caused by the fact that the highest bids in that region are very rarely observed, disappears with smaller N and/or sufficiently large n.

Table 1.1 presents the implied bounds on the optimal reserve prices for different parameters of the Log-Normal distribution. While the bounds estimated with both methods are fairly wide, the LAM estimates are, on average, substantially tighter.

Figure 1.4: Estimated Bounds on the CDF of Valuations



Note: The number of bidders is $N = 6$, the sample size is $n = 200$. The dashed lines represent the average estimates for the bounds across simulations. The lower bound is the same for both estimation methods. The dotted lines represent the 5-th and 95-th quantiles across simulations.

1.8 Extension to Convex Cones

In the settings where the tangent set $T(P)$ is a convex cone, the lower bound in Theorem 1.1 holds with $S(\mathbb{G}_0)$ replaced by $\theta_0'(T(P))$. Such settings typically arise in the presence of moment inequality restrictions that are binding at P. Common examples include point- or over-identifying moment inequality models, or regression models with binding sign constraints. However, such settings are theoretically problematic: when $T(P)$ is a convex cone, the optimal estimators proposed by Convolution and Minimax Theorems may often be inadmissible, even for differentiable parameters.[22] To illustrate, I consider a simple example, similar to Imbens and Manski (2004).

Suppose that the parameter of interest $\theta_0 \in \mathbb{R}$ is partially identified, and the bounds are given by $\theta_{L,0} = \theta_L(P)$ and $\theta_{U,0} = \theta_U(P)$, which are "smooth" functionals (i.e., differentiable in the sense of Definition 1.4) of the distribution P of the observable random vector X. The model is given by:[23]

$$\mathbf{P} = \{P : \theta_L(P) \leqslant \theta_U(P)\}$$

What is an efficient estimator for the identified set $[\theta_{L,0}, \theta_{U,0}]$? In this example, stimating the identified set amounts to estimating a two-dimensional vector of bounds. First, consider a situation when $\theta_L(P) < \theta_U(P)$. In this case, the tangent set is unrestricted, i.e., $T(P) = L_2^0(P)$, and the classical efficiency theory suggests that the "plug-in" estimator, defined by $\hat{\theta}_{L,n} \equiv \theta_L(\hat{P}_n)$ and $\hat{\theta}_{U,n} \equiv \theta_U(\hat{P}_n)$, where $\hat{P}_n$ denotes the empirical distribution, is optimal. Intuitively, the bounds can be estimated separately because they are not informative about each other. On the other hand, suppose that $\theta_L(P) = \theta_U(P)$. In this case, the estimators $\hat{\theta}_{L,n}$ and $\hat{\theta}_{U,n}$ target the same parameter, so the intuition suggests that they may be combined to produce a more efficient estimator. For example, assuming that the asymptotic variances of

[22] More specifically, if $T(P)$ is a cone but $\overline{\text{lin}}\ T(P) = L_2^0(P)$, the optimal estimator suggested by the Convolution and Minimax Theorems will be the same as the estimator when $T(P) = L_2^0(P)$, e.g. van der Vaart (1988).

[23] The model may be required to satisfy some other restrictions omitted here for simplicity.

$\hat{\theta}_{L,n}$ and $\hat{\theta}_{U,n}$ are the same, the optimal GMM would suggest using $(\hat{\theta}_{L,n}+\hat{\theta}_{U,n})/2$ to estimate both θ_L and θ_U. However, due to the tangent set being a cone, the existing semiparametric efficiency theory suggests otherwise. More precisely, denoting the path-wise derivatives by $\theta'_{0,L}(h) = \mathbb{E}_P(\psi_L h)$ and $\theta'_{0,U}(h) = \mathbb{E}_P(\psi_U h)$ for some $\psi_L, \psi_U \in L_2^0(P)$, the tangent set is given by

$$T(P) = \{h \in L_2^0(P) : \mathbb{E}_P((\psi_L(X) - \psi_U(X))h(X)) \leqslant 0\}$$

Then, since $\overline{\text{lin}}\ T(P) = L_2^0(P)$, both the Convolution Theorem and LAM Theorem suggest that the "plug-in" estimator $[\hat{\theta}_{L,n}, \hat{\theta}_{U,n}]$ is still optimal, which contradicts the above intuition.[24]

The above example shows that the existing semiparametric efficiency theory cannot properly capture binding inequality constraints. Although dealing with such inconsistency is beyond the scope of this paper, it is an interesting question for further research.

1.9 Conclusion

In many econometric models, certain parameters of interest are represented via directionally differentiable functionals. The potential lack of full differentiability has raised concerns in regard to choosing "good" estimators for such parameters. This paper proposed a solution by deriving Locally Asymptotically Minimax estimators within a class of plug-in estimators with additive adjustment terms. In contrast with fully differentiable settings, the optimal estimator depends on the chosen loss function, suggesting that it must be tailored to specific applications. The proposed estimators typically do not reduce the bias as much as some of the existing methods, but avoid large fluctuations in risk around the points where differentiability fails. Empirical relevance of the proposed method was demonstrated in an application to English auctions with independent private values.

[24]The Convolution Theorem continues to hold under the assumption that $T(P)$ is a convex cone if formulated with $\overline{\text{lin}}\, T(P)$ instead of $T(P)$.

1.10 Appendix: Proofs from the Main Text

1.10.1 Known Results for Reference

The following results refer to Definitions 1.4 and 1.5.

Theorem 1.3 (Convolution Theorem for Euclidean Parameters. Theorem 25.20 in van der Vaart (2000))**.** *Assume that $\theta(P) \in \mathbb{R}^{d_\theta}$ is differentiable relative to a tangent set $T(P)$ with the path-wise derivative $\theta_0' : \bar{T}(P) \to \mathbb{R}^d$. Then, for any regular estimator sequence $\hat{\theta}_n$,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{P_{n,0}}{\rightsquigarrow} Z + W,$$

where Z is a centered Gaussian random vector in $\mathbb{R}^{d_\theta}$, and W is a tight random vector in $\mathbb{R}^{d_\theta}$ independent from Z. The covariance matrix of Z is given by $\Sigma = \mathbb{E}(\tilde{\theta}\tilde{\theta}^T)$, where $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{d_\theta})^T$ is the efficient influence function for $\theta(P)$. That is, $\tilde{\theta}_j \in T(P)$, for $j = 1, \ldots, d_\theta$, are such that $\theta_0'(h) = \mathbb{E}_P(\tilde{\theta}h)$ for all $h \in T(P)$. Moreover, the distribution of Z concentrates on the range of Σ.

To state the Convolution Theorem for infinite-dimensional parameters, some new notation is required. For each $b^* \in \mathbb{B}^*$ (the continuous dual of $\mathbb{B}$), $b^* \circ \theta_0'$ is a continuous linear map from $\bar{T}(P)$ into $\mathbb{R}$. By the Riesz Representation Theorem (Theorem **??**), there is an element $\tilde{\theta}_{b^*} \in \bar{T}(P)$ such that $b^* \circ \theta_0'(h) = \mathbb{E}_P(\tilde{\theta}_{b^*}h)$ for any $h \in \bar{T}(P)$. Such $\tilde{\theta}_{b^*}$ is called the *canonical gradient* of θ in the direction b^*.

Theorem 1.4 (Convolution Theorem. Theorem 3.11.2. in van der Vaart and Wellner (1996))**.** *Assume that $\theta(P) \in \mathbb{B}$ is differentiable relative to a tangent set $T(P)$ with the path-wise derivative $\theta_0' : \bar{T}(P) \to \mathbb{B}$. Then, for any regular estimator sequence $\hat{\theta}_n$,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{P_{n,0}}{\rightsquigarrow} \mathbb{G}_0 + \mathbb{W},$$

where $\mathbb{G}_0$ is a tight centered Gaussian random element in $\mathbb{B}$ and $\mathbb{W}$ is a tight random element in $\mathbb{B}$ independent from $\mathbb{G}_0$. The distribution of $\mathbb{G}_0$ is such that $(b_1^*, \ldots, b_K^*) \circ \mathbb{G}_0$ is a cen-

tered Gaussian random vector with $\mathbb{C}ov(b_i^*(\mathbb{G}_0), b_j^*(\mathbb{G}_0)) = \mathbb{E}(\tilde{\theta}_{b_i^*}\tilde{\theta}_{b_j^*})$ *for any* $b_1^* \dots b_K^* \in \mathbb{B}^*$. *Moreover, the distribution of* $\mathbb{G}_0$ *concentrates on the closure of* $\theta_0'(T(P))$.

Theorem 1.5 (Continuous Mapping Theorem. Theorem 1.3.6. in van der Vaart and Wellner (1996))**.** *Let a map between two metric spaces* $g : \mathbb{B} \to \mathbb{D}$ *be continuous at every point of a set* $\mathbb{B}_0 \subset \mathbb{B}$. *If* $X_n \rightsquigarrow X$ *and* X *takes its values in* $\mathbb{B}_0$, *then* $g(X_n) \rightsquigarrow g(X)$.

Theorem 1.6 (Prohorov's Theorem. Theorem 1.3.9. in van der Vaart and Wellner (1996))**.** *If the sequence* X_n *is asymptotically tight and asymptotically measurable, then for any subsequence* $X_{n'}$ *there is a further sibsequence* $X_{n''}$ *that converges weakly to a tight Borel law.*

Let (X, ρ) denote a metric space and $B \subset X$ be an arbitrary subset of X. For each $x \in X$ define $\rho(x, B) = \inf\{\rho(x, y)|y \in B\}$, which may be infinite.

Lemma 1.3 (Suprema of Lower Semi-Continuous Functions In Polish Spaces)**.**
Let (X, ρ) *be a separable metric space,* $B \subset X$ *be an arbitrary non-empty subset and* $f : X \to \mathbb{R}$ *be a lower semi-continuous function. Then* B *is separable and*

$$\sup_B f(x) = \sup_{B^\circ} f(x),$$

where B° *denotes a countable dense subset of* B.

Proof. First, I show that B is separable. Let $E = \{e_1, e_2, \dots\}$ denote a countable dense subset of X. Fix $\varepsilon > 0$. Define $E' = \{e_j \in E|\rho(e_j, B) \leqslant \varepsilon/3\} = \{e_1', e_2', \dots\}$ which is non-empty since E is dense in X. For every such $e_j' \in E'$ there is $x_j \in B$ with $\rho(e_j', x_j) \leqslant \rho(e_j', B) + \varepsilon/3 \leqslant 2\varepsilon/3$. Let B° denote a set of all $x_j \in B$ obtained this way. Since E is dense in X, for any $x \in B$ there is $e_k \in E$ with $\rho(e_k, x) \leqslant \varepsilon/3$. Since $\rho(e_k, B) \leqslant \rho(e_k, x)$ by definition, it must be that $e_k = e_j'$ for some $e_j' \in E'$ and $\rho(e_j', x) \leqslant \varepsilon/3$. For such e_j' there is $x_j \in B^\circ$ with $\rho(e_j', x_j) \leqslant 2\varepsilon/3$. By triangle inequality, $\rho(x, x_j) \leqslant \rho(x, e_j') + \rho(e_j', x_j) \leqslant \varepsilon$ so that B° is a countable dense subset of B.

For the second part of the statement, it is clear that $\sup_{B^\circ} f(x) \leqslant \sup_B f(x)$. For the reversed inequality, it suffices to show $\sup_B f(x) \leqslant \sup_{B^\circ} f(x) + \varepsilon$ for an arbitrary $\varepsilon > 0$.

Pick $x' \in B$ such that $\sup_B f(x) \leqslant f(x') + \varepsilon$. Since B° is dense in B, there is a sequence $(x_n)_{n\geqslant 1} \in B^\circ$ such that $\rho(x_n, x') \to 0$. It follows from lower semi-continuity of f that $\liminf_{n\to\infty} f(x_n) \geqslant f(x')$. Therefore, $\sup_B f(x) \leqslant \liminf_{n\to\infty} f(x_n) + \varepsilon \leqslant \sup_{B^\circ} f(x) + \varepsilon$, and the proof is complete.

■

Lemma 1.4 (Uniform Convergence of Lipchitz Functions)**.** *Let (X, ρ) denote a compact metric space and $f_n : X \to \mathbb{R}$ be a uniformly Lipchitz sequence of functions, that is, for some constant C independent of n,*

$$|f_n(x) - f_n(x')| \leqslant C \cdot \rho(x, x').$$

If $f_n(x)$ converges point-wise ti some $f : X \to \mathbb{R}$, then f is Lipchitz with the same constant and $\sup_{x\in X} |f_n(x) - f(x)| \to 0$.

Proof. First, I show that f satisfies:

$$|f(x) - f(x')| \leqslant C\rho(x, x')$$

for any $x, x' \in K$. Fix $\delta > 0$. Choose n_1 and n_2 such that $|f_n(x) - f(x)| < \delta$ for all $n \geqslant n_1$ and $|f_n(x') - f(x')| < \delta$ for all $n \geqslant n_2$. Then, for any $n \geqslant \max\{n_1, n_2\}$,

$$|f(x) - f(x')| \leqslant |f(x) - f_n(x)| + |f_n(x) - f_n(x')| + |f_n(x') - f(x')| \leqslant C\rho(x, x') + 2\delta.$$

Since δ was arbitrary, the desired conclusion follows.

Next, fix some $\varepsilon > 0$. Since K is compact, there are $x_1, \ldots, x_J$ such that $K \subset \bigcup_{j=1}^J B(x_j, \varepsilon)$. Let $\pi : K \to \{x_1, \ldots, x_J\}$ be defined by $\pi(x) = \operatorname{argmin}_{j\leqslant j}\{\rho(x, x_j)\}$, so that $\rho(x, \pi x) \leqslant \varepsilon$ for any $x \in X$. Then

$$\begin{aligned} \sup_{x\in K} |f_n(x) - f(x)| \quad &\leqslant \sup_{x\in K} |f_n(x) - f_n(\pi x)| \qquad (I) \\ &+ \sup_{x\in K} |f_n(\pi x) - f(\pi x)| \qquad (II) \\ &+ \sup_{x\in K} |f(\pi x) - f(x)|. \qquad (III) \end{aligned}$$

Note that $(I) \leqslant C\varepsilon$ and $(III) \leqslant C\varepsilon$ by construction, and $(II) = \max_{j\leqslant J} |f_n(x_j) - f(x_j)| = o(1)$. Letting $n \to \infty$ followed by $\varepsilon \to 0$ concludes the proof. ■

1.10.2 Auxiliary Lemmas

Lemma 1.5 (Lipchitzness of the Asymptotic Risk)**.** *Let l_M be a loss function satisfying Remark 1.1, and ϕ be a directionally differentiable function satisfying Assumption 1.2.1. Let $(\mathbb{D}, ||\cdot||_{\mathbb{D}})$ and $(\mathbb{B}, ||\cdot||_{\mathbb{B}})$ denote Banach spaces and Z denote a tight random element in $\mathbb{B}$. Then a function $f : \mathbb{D} \times \mathbb{B} \times \mathbb{B} \to \mathbb{R}$ defined as*

$$f(v, w, r) = \mathbb{E}\left(l_M(v - \phi_0'(Z + w + r) + \phi_0'(r))\right)$$

is jointly Lipchitz, i.e. $|f(v, w, r) - f(\tilde{v}, \tilde{w}, \tilde{r})| \leqslant C_{M,\phi} \cdot (||v - \tilde{v}||_{\mathbb{D}} + ||w - \tilde{w}||_{\mathbb{B}} + ||r - \tilde{r}||_{\mathbb{B}})$ for all (v, w, r), and $(\tilde{v}, \tilde{w}, \tilde{r})$, for some $C_{M,\phi} < \infty$.

Proof. Let $\Delta f = f(v, w, r) - f(\tilde{v}, \tilde{w}, \tilde{r})$ and $C_{M,\phi} = \max(C_M, 2C_M C_\phi)$. By Jensen's inequality, the assumed Lipchitzness of l_M and ϕ_0', and triangle inequality:

$$\begin{aligned}
|\Delta f| &\leqslant \mathbb{E}\left(|l_M(v - \phi_0'(Z + w + r) + \phi_0'(r)) - l_M(\tilde{v} - \phi_0'(Z + \tilde{w} + \tilde{r}) + \phi_0'(\tilde{r}))|\right) \\
&\leqslant C_M\left(||v - \tilde{v}||_{\mathbb{D}} + ||\phi_0'(r) - \phi_0'(\tilde{r})||_{\mathbb{D}} + \mathbb{E}\left(||\phi_0'(Z + w + r) - \phi_0'(Z + \tilde{w} + \tilde{r})||_{\mathbb{D}}\right)\right) \\
&\leqslant C_M \cdot \left(||v - \tilde{v}||_{\mathbb{D}} + C_\phi\, ||r - \tilde{r}||_{\mathbb{B}} + C_\phi(||w - \tilde{w}||_{\mathbb{B}} + ||r - \tilde{r}||_{\mathbb{B}}) \right) \\
&\leqslant C_{M,\phi} \cdot (||v - \tilde{v}||_{\mathbb{D}} + ||w - \tilde{w}||_{\mathbb{B}} + ||r - \tilde{r}||_{\mathbb{B}})\,.
\end{aligned}$$

■

Lemma 1.6 (Approximating Sub-Convex Loss Functions)**.** *Any subconvex loss function l (see Assumption 1.3.3) can be approximated by a sequence of bounded Lipschitz functions l_M pointwise monotonically from below.*

Proof. First, note that the sequence of bounded step functions $\{l_r\}$ defined as

$$l_r(x) = \frac{1}{2^r} \sum_{i=1}^{2^{2r}} \mathbf{1}\left\{x : l(x) > \frac{i}{2^r}\right\} = \sum_{i=1}^{2^{2r}} \frac{i}{2^r} \cdot \mathbf{1}\left\{x : \frac{i}{2^r} < l(x) \leqslant \frac{i+1}{2^r}\right\}$$

converges to l pointwise monotonically from below. Next, introduce the sets $A_i = \left\{x : \frac{i}{2^r} < l(x) \leqslant \frac{i+1}{2^r}\right\}$ and $B_i = \cup_{j \leqslant i} A_j$ and let $F_{M,i} = \{x \in A_i : d(x, B_i) \geqslant 1/M\}$. For a fixed r, consider a sequence of functions, $\{l_{M,r}\}$, defined as

$$l_{M,r}(x) = \sum_{i=1}^{2^{2r}} \left(\frac{i-1}{2^r} + \frac{d(x, B_i)}{d(x, B_i) + d(x, F_{M,i})} \right) \cdot \mathbf{1}(x \in A_i)$$

Every such function is bounded by 2^r and the part $d(x, B_i)/(d(x, B_i) + d(x, F_{M,i}))$ smoothes out the jumps in l_r, such that the resulting function is Lipschitz continuous with Lipschitz constant equal to $M/2^r$. Indeed, let $y \in A_j$, $x \in A_i$ with $j \geqslant i$

$$l_{M,r}(y) - l_{M,r}(x) = \frac{j-i}{2^r} + \frac{1}{2^r} \left(\frac{d(y, B_j)}{d(y, B_j) + d(y, F_{M,j})} - \frac{d(x, B_i)}{d(x, B_i) + d(x, F_{M,i})} \right)$$

First, let $i = j$. Then

$$\begin{aligned}
|l_{M,r}(y) - l_{M,r}(x)| &= \frac{1}{2^r} \left| \frac{d(y, B_i) d(x, F_{M,i}) - d(x, B_i) d(y, F_{M,i})}{(d(y, B_i) + d(y, F_{M,i}))(d(x, B_i) + d(x, F_{M,i}))} \right| \\
&= \frac{1}{2^r} \left| \frac{d(y, B_i)(d(x, F_{M,i}) - d(y, F_{M,i})) + d(y, F_{M,i})(d(y, B_i) - d(x, B_i))}{(d(y, B_i) + d(y, F_{M,i}))(d(x, B_i) + d(x, F_{M,i}))} \right| \\
&\overset{(a)}{\leqslant} \frac{1}{2^r} \cdot \frac{(d(y, B_i) + d(y, F_{M,i})) \cdot d(x, y)}{(d(y, B_i) + d(y, F_{M,i}))(d(x, B_i) + d(x, F_{M,i}))} \\
&\overset{(b)}{\leqslant} \frac{M}{2^r} \cdot d(x, y)
\end{aligned} \tag{1.29}$$

Where (a) follows from the reverse triangle inequality, i.e. $|d(y, B_i) - d(x, B_i)| \leqslant d(x, y)$ and similar for $F_{M,i}$, and (b) follows from the fact that $d(x, B_i) + d(x, F_{M,i}) \geqslant 1/M$ by construction. The same upper bound can be obtained in a straightforward way when $j \geqslant i + 1$ by considering four different cases when $y \in F_{M,j}$ or $y \in A_j \backslash F_{M,j}$ and $x \in F_{M,i}$ or $x \in A_i \backslash F_{M,i}$. ■

Lemma 1.7 (Point-wise Consistency of Set Extremum Estimators)**.** *Let $(\mathcal{V}, d)$ be a metric space. Let $\hat{Q}_n(v)$ and $Q(v)$ denote the empirical and population criterion functions, correspondingly. Let $\mathcal{V}_0$ denote the set of maximizers of the population criterion function and $\hat{v}_n$*

denote any "almost maximizer" of $\hat{Q}_n$ over a sieve space $\mathcal{V}_{k(n)}$, i.e.

$$\hat{Q}_n(\hat{v}_n) \geqslant \sup_{v \in \mathcal{V}_{k(n)}} \hat{Q}_n(v) - O_P(\eta_{k(n)})$$

Assume that the following conditions hold.

1. *(Identification) For each $v_0 \in \mathcal{V}_0$:*

$$Q(v_0) - \sup_{\{v \in \mathcal{V}_k \colon d(v, \mathcal{V}_0) \geqslant \varepsilon\}} Q(v) > \delta(k) \cdot g(\varepsilon) \quad \text{for all } k \geqslant 1 \text{ and } \varepsilon > 0$$

   for a positive non-increasing function $\delta(k)$ and positive $g(\varepsilon)$.

2. *(Sieve Approximation) The sieve spaces $\mathcal{V}_k \subset \mathcal{V}_{k+1} \subset \ldots$ are compact under d and grow dense in $\mathcal{V}$ in a sense that there is a sequence of maps $\pi_k : \mathcal{V} \to \mathcal{V}_k$ such that for each $v_0 \in \mathcal{V}_0$ it holds that $d(v_0, \pi_k v_0) \to 0$ as $k \to \infty$.*

3. *(Continuity) $Q(v)$ is upper semi-continuous on all $\mathcal{V}_k$ with $|Q(v_0) - Q(\pi_k v_0)| = o(\delta(k))$ for each $v_0 \in \mathcal{V}_0$.*

4. *(Uniform Convergence and Quality of Maximization)*

   (a) for each fixed $k \geqslant 1$: $\sup_{v \in \mathcal{V}_k} |\hat{Q}_n(v) - Q(v)| = o_P(1)$ as $n \to \infty$

   (b) $\sup_{v \in \mathcal{V}_{k(n)}} \left|\hat{Q}_n(v) - Q(v)\right| \equiv \hat{c}_{k,n} = o_P(\delta(k(n)))$

   (c) $\eta_{k(n)} = o(\delta(k(n)))$

Let $\hat{V}_n$ denote the set of "almost maximizers" of $\hat{Q}_n$. Then, $\vec{d}_H(\hat{V}_n, \mathcal{V}_0) = o_P(1)$, where $\vec{d}_H(A, B) = \sup_{a \in A} \inf_{b \in B} d(a, b)$ denotes the directed Hausdorff distance.

Proof. Let $(\Omega_n, \mathcal{A}_n, P_n)$ denote a sequence of probability spaces. The maps $\hat{Q}_n(v) : \Omega_n \to \mathbb{R}$ are not required to be measurable, and, throughout the proof, the "events" defined via $\hat{Q}_n$ are thought of as subsets of Ω_n rather than elements of $\mathcal{A}_n$, and all probabilities are outer probabilities.

Some familiar properties of probability hold for outer probability as well. In particular, let $A, B, C, D \subset \Omega_n$. Then for $A \subset B$ it holds that $P^*(A) \leqslant P^*(B)$, and if $C \cap D = \varnothing$, it holds that $P^*(C \cup D) \leqslant P^*(C) + P^*(D)$. See Lemmas 1.2.2 and 1.2.3 in van der Vaart and Wellner (1996) for the details.

Notice that $d(\hat{v}_n, \mathcal{V}_0) \geqslant \varepsilon$ implies that $\hat{Q}_n$ is almost-maximized (at $\hat{v}_n$) at least ε-away from $\mathcal{V}_0$. Let $\mathcal{V}^{\varepsilon}_{k(n)} = \{v \in \mathcal{V}_{k(n)} : d(v, \mathcal{V}_0) \geqslant \varepsilon\}$, which, by Condition 2, is a compact set. Therefore,

$$\begin{aligned} P(d(\hat{v}_n, \mathcal{V}_0) \geqslant \varepsilon) &\leqslant P\left(\sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} \hat{Q}_n(v) \geqslant \sup_{v \in \mathcal{V}_{k(n)}} \hat{Q}_n(v) - O_P(\eta_{k(n)}) \right) \\ &\leqslant P\left(\sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} \hat{Q}_n(v) \geqslant \hat{Q}_n(\pi_{k(n)} v_0) - O_P(\eta_{k(n)}) \right) \end{aligned}$$

where the second inequality is valid for all $v_0 \in \mathcal{V}_0$. Call the latter event A_n and write is as:

$$\begin{aligned} A_n = \Bigg\{ &\sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} Q(v) - Q(\pi_{k(n)} v_0) + O_P(\eta_{k(n)}) \\ &\geqslant \hat{Q}_n(\pi_k v_0) - Q(\pi_{k(n)} v_0) + \sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} Q(v) - \sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} \hat{Q}_n(v) \Bigg\} \end{aligned}$$

Consider a sequence of events $(B_n)_{n \geqslant 1}$ defined as

$$B_n = \left\{ \sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} \left| \hat{Q}_n(v) - Q(v) \right| > \hat{w}_{k(n)} \right\}$$

for some sequence $\hat{w}_{k(n)}$ to be chosen later. Note that B_n^c implies

$$\left\{ \left| \sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} \hat{Q}_n(v) - \sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} Q(v) \right| \leqslant \hat{w}_{k(n)} \right\} \implies \begin{cases} \sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} Q(v) \geqslant \sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} \hat{Q}_n(v) - \hat{w}_{k(n)} \\ \hat{Q}_n(\pi_{k(n)} v_0) \geqslant Q(\pi_{k(n)} v_0) - \hat{w}_{k(n)} \end{cases}$$

With the above notation, write $P(A_n) \leqslant P(B_n) + P(A_n \cap B_n^c)$ to obtain:

$$\begin{aligned} P(A_n) &\leqslant P(B_n) + P\left(\sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} Q(v) - Q(\pi_{k(n)} v_0) + O_P(\eta_k) \geqslant -2\hat{w}_{k(n)} \right) \\ &\leqslant P(B_n) + P\left(2\hat{w}_{k(n)} + O_P(\eta_{k(n)}) + |Q(v_0) - Q(\pi_{k(n)} v_0)| \geqslant Q(v_0) - \sup_{v \in \mathcal{V}^{\varepsilon}_{k(n)}} Q(v) \right) \end{aligned}$$

Consider, specifically, $\hat{w}_{k(n)} = \hat{c}_{k,n} = o_P(\delta(k(n)))$. Then $P(B_n) = 0$ by the definition of $\hat{c}_{k,n}$ in Condition 4, and the second probability converges to zero by the choice of $\hat{w}_{k(n)}$ and Conditions 1, 3 and 4. Since the upper bound does not depend on the choice of $\hat{v}_n \in \hat{V}_n$, it follows that $\vec{d}_H(\hat{V}_n, \mathcal{V}_0) = o_P(1)$.

■

Lemma 1.8 (Replacing The Feasible Set)**.** *Let $(\mathbb{B}, ||\cdot||_{\mathbb{B}})$ be a Banach space, $K \in \mathbb{B}$ be a compact set and $f_n : \mathbb{B} \times \mathbb{B} \to \mathbb{R}$ be a sequence of random functions satisfying, for each $x_1, x_2 \in \mathbb{B}$,*

$$\sup_{v \in K} |f_n(x_1; v) - f_n(x_2; v)| \leqslant C_n \cdot ||x_1 - x_2||_{\mathbb{B}}$$

for a possibly random positive sequence $C_n = O_P(1)$. Further, let $(\hat{A}_n)_{n \geqslant 1}$ and $(A_n)_{n \geqslant 1}$ denote sequences of measurable sets in $\mathbb{B}$ such that $\sup_{x \in \hat{A}_n} f_n(x; v)$ and $\sup_{x \in A_n} f_n(x; v)$ are attained at some points for each n. If $d_H(\hat{A}_n, A_n) = o_P(1)$, then:

$$\sup_{v \in K} \left| \sup_{x \in \hat{A}_n} f_n(x; v) - \sup_{x \in A_n} f_n(x; v) \right| = o_P(1)$$

Proof. Let $\hat{\Delta}_n$ denote the left-hand side of the preceding display and take any $\hat{x}_n$ and x_n that attain the suprema of f over $\hat{A}_n$ and A_n correspondingly. By assumption, for each $\varepsilon > 0$, $||x_1 - x_2||_{\mathbb{B}} < \delta_n$ implies $\sup_{v \in K} |f_n(x_1; v) - f_n(x_2; v)| < \varepsilon$ where $\delta_n = \varepsilon / C_n$.

Note that $d_H(\hat{A}_n, A_n) < \delta_n$ implies that (1) for $\hat{x}_n \in \hat{A}_n$, there is $\tilde{x}_n \in A_n$ with $||\hat{x}_n - \tilde{x}_n||_{\mathbb{B}} < \delta_n$ and (2) for $x_n \in A_n$, there is $x'_n \in \hat{A}_n$ with $||x_n - x'_n||_{\mathbb{B}} < \delta_n$. Then, by Lipschitz continuity of f_n, for each v it holds that (1) $f_n(\tilde{x}_n; v) > f_n(\hat{x}_n; v) - \varepsilon$ and therefore $f_n(x_n; v) > f_n(\hat{x}_n; v) - \varepsilon$ and (2) $f_n(x'_n; v) > f_n(x_n; v) - \varepsilon$ and therefore $f_n(\hat{x}_n; v) > f_n(x_n; v) - \varepsilon$. These inequalities combined give $\sup_{v \in K} |f_n(\hat{x}_n; v) - f_n(x_n; v)| = \hat{\Delta}_n < \varepsilon$. Therefore, taking contrapositive,

$$P(\hat{\Delta}_n > \varepsilon) = P(\sup_{v \in K} |f_n(\hat{x}_n; v) - f_n(x_n; v)| > \varepsilon) \leqslant P(d_H(\hat{A}_n, A_n) > \delta_n) \to 0$$

as $n \to \infty$, which completes the proof.

■

1.10.3 Proof of Theorem 1.1

Consider an estimator sequence of the form

$$\hat{\phi}_n = \phi\left(\hat{\theta}_n + \frac{\hat{v}_{1,n}}{\sqrt{n}}\right) + \frac{\hat{v}_{2,n}}{\sqrt{n}}, \tag{1.30}$$

where $\hat{\theta}_n$ is the best regular estimator for θ_0 in the sense of the Convolution Theorem (1.4), and $\hat{v}_{1,n}$, $\hat{v}_{2,n}$ are adjustment terms depending on the data. To calculate the LAM risk of this estimator sequence, it is necessary to study its distributional limits under the "local perturbations" $P_{n,h}$ (see Definition 1.5).

Let $v_1 \in \mathbb{B}$ and $v_2 \in \mathbb{D}$ denote the probability limits of $\hat{v}_{1,n}$ and $\hat{v}_{2,n}$ under $P_{n,h}$ correspondingly, which are the same as under $P_{n,h}$ by contiguity; see Lemma 6.4 in van der Vaart (2000). Since $\hat{\theta}_n$ is the best regular estimator, $\sqrt{n}(\hat{\theta}_n - \theta(P_{n,h})) \rightsquigarrow_{P_{n,h}} \mathbb{G}_0$ for all $h \in T(P)$. Since $\theta(P)$ is differentiable, $\sqrt{n}(\theta(P_{n,h}) - \theta_0) = \theta_0'(h)$ for any $h \in T(P)$. By the Prohorov's Theorem, for any subsequence, there is a further subsequence, still denoted by n for simplicity, such that:

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n + \tfrac{\hat{v}_{1,n}}{\sqrt{n}} - \theta_0\right) &= \sqrt{n}\left(\hat{\theta}_n - \theta(P_{n,h})\right) + \hat{v}_{1,n} + \sqrt{n}(\theta(P_{n,h}) - \theta_0) \\ &\overset{P_{n,h}}{\rightsquigarrow} \mathbb{G}_0 + v_1 + \theta_0'(h)\end{aligned}$$

as random elements in $\mathbb{B}$. The assumed differentiability of $\theta(P)$ allows to write $\theta(P_{n,h}) = \theta_0 + \theta_0'(h)/\sqrt{n} + o(1/\sqrt{n})$, in $\mathbb{B}$. By the directional differentiability of ϕ and Delta-method for directionally differentiable functions,[25]

$$\begin{aligned}\sqrt{n}\left(\hat{\phi}_n - \phi(\theta(P_{n,h}))\right) &= \sqrt{n}(\phi(\hat{\theta}_n + \tfrac{\hat{v}_{1,n}}{\sqrt{n}}) - \phi(\theta_0)) - \sqrt{n}(\phi(\theta(P_{n,h})) - \phi(\theta_0)) + \hat{v}_{2,n} \\ &\overset{P_{n,h}}{\rightsquigarrow} \phi_0'(\mathbb{G}_0 + v_1 + \theta_0'(h)) - \phi_0'(\theta_0'(h)) + v_2\end{aligned}$$

as random elements in $\mathbb{D}$. Let $l_M \leqslant l$ be a loss function satisfying Remark 1.1. Then, by the Portmanteau Theorem,

$$\mathbb{E}_{P_{n,h}}\left\{l_M\left(\sqrt{n}\left(\hat{\phi}_n - \phi(\theta(P_{n,h}))\right)\right)\right\} \to \mathbb{E}\left\{l_M\left(\phi_0'(\mathbb{G}_0 + v_1 + \theta_0'(h)) - \phi_0'(\theta_0'(h)) + v_2\right)\right\}$$

[25] If $\sqrt{n}(\hat{\gamma}_n - \gamma_0) \rightsquigarrow Z$ and f is Hadamard directionally differentiable at γ_0 with directional derivative f', then $\sqrt{n}(f(\hat{\gamma}_n) - f(\gamma_0)) \rightsquigarrow f'(Z)$. See Shapiro (1990).

as $n \to \infty$, uniformly in h in any finite set $I \subset T(P)$. Therefore,

$$\sup_{I \subset T(P)} \liminf_{n\to\infty} \sup_{h \in I} \mathbb{E}_{P_{n,h}} \left\{ l\left(\sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h)))\right)\right\}$$
$$\geqslant \sup_{h \in T(P)} \mathbb{E}\left\{ l_M\left(\phi_0'(\mathbb{G}_0 + v_1 + \theta_0'(h)) - \phi_0'(\theta_0'(h)) + v_2\right)\right\}. \quad (1.31)$$

The supremum in the second line of the above display can be equivalently taken over $s \in \theta_0'(T(P))$ and further over the closure of this set in $\mathbb{B}$ (by Lemma 1.3), which is equal to $S(\mathbb{G}_0)$. Then, the result follows by passing to a limit as $M \to \infty$, invoking the Monotone Convergence Theorem, and then taking an infimum with respect to all v_1, v_2.

1.10.4 Proof of Theorem 1.2

I will show that $\hat{v}_{1,n}, \hat{v}_{2,n}$ converge in probability (along subseqences) to some minimizers of the lower bound. In view of Lemma 1.7, it suffices to show that Assumptions 1 (identification condition) and 4 (uniform convergence) there are satisfied. Since K is compact and the criterion function is continuous, the identification condition is immediate, so I will show the uniform convergence. Denote:

$$\hat{g}_n(b, v, s) = l_M(\hat{\phi}_n'(b + v_1 + s) - \hat{\phi}_n(s) + v_2),$$

$$g(b, v, s) = l_M(\phi_0'(b + v_1 + s) - \phi_0'(s) + v_2).$$

Note that for any $v \in K, b \in \mathbb{B}, c \in \mathbb{B}$

$$|\hat{g}_n(b, v, s) - g(b, v, s)|$$
$$\leqslant C_M \left(\left|\left| \hat{\phi}'(b + v_1 + s) - \phi_0'(b + v_1 + s) \right|\right| + \left|\left| \hat{\phi}_n'(s) - \phi_0'(s) \right|\right| \right). \quad (1.32)$$

Let:

$$\hat{Q}_{1,n}(v) = \sup_{s \in \hat{R}_n} \mathbb{E}\left(\hat{g}_n\left(\hat{\mathbb{G}}_n^*, v, s\right) \middle| X_1^n \right),$$
$$\hat{Q}_{2,n}(v) = \sup_{s \in R_n} \mathbb{E}\left(\hat{g}_n\left(\hat{\mathbb{G}}_n^*, v, s\right) \middle| X_1^n \right),$$

$$\hat{Q}_{3,n}(v) = \sup_{s \in R_n} \mathbb{E}\left(g\left(\hat{\mathbb{G}}_n^*, v, s\right) \middle| X_1^n \right),$$
$$Q_{4,n}(v) = \sup_{s \in R_n} \mathbb{E}\left(g\left(\mathbb{G}_0, v, s\right)\right),$$
$$Q(v) = \sup_{s \in S(\mathbb{G}_0)} \mathbb{E}\left(g(\mathbb{G}_0, v, s)\right).$$

First, $\sup_{v \in K} |\hat{Q}_{1,n}(v) - \hat{Q}_{2,n}(v)| = o_P(1)$ follows immediately from Lemma 1.8 and the fact that $d_H(\hat{R}_n, R_n) = o_P(1)$ by Assumption 1.5.3. To show that Lemma 1.8 can be applied with $f_n(s; v) = \mathbb{E}(\hat{g}_n(\hat{\mathbb{G}}_n^*, v, s) \mid X_1^n)$, note that

$$\sup_{v \in K} |f_n(s_1; v) - f_n(s_2; v)| \leqslant C_M \cdot C_{\hat{\phi}_n'} \cdot ||s_1 - s_2||.$$

Second, $\sup_{v \in K} |\hat{Q}_{2,n}(v) - \hat{Q}_{3,n}(v)| = o_P(1)$ follows from the assumed uniform consistency of $\hat{\phi}_n'$ in Assumption 1.5.2. Indeed, note that Assumption 1.5.1 implies that $\hat{\mathbb{G}}_n^*$ converges weakly to $\mathbb{G}_0$ unconditionally (see Lemma S.3.1. in the supplemental appendix to Fang and Santos, 2019). Next, fix any $\varepsilon > 0$ and $\eta > 0$. Since $\mathbb{G}_0$ is tight, there is a compact set $S \subset \mathbb{B}$ such that $P(\mathbb{G}_0 \notin S) \leqslant \varepsilon\eta$. Then, by the Portmanteau Theorem, for any $\delta > 0$

$$\limsup_{n \to \infty} P(\hat{\mathbb{G}}_n^* \notin S^\delta) \leqslant P(\mathbb{G}_0 \notin S) \leqslant \varepsilon\eta$$

Therefore, by Markov's inequality and Fubini's Theorem (Lemma 1.2.6. in van der Vaart and Wellner, 1996),

$$\limsup_{n \to \infty} P(P(\hat{\mathbb{G}}_n^* \notin S^\delta | X_1^n) > \eta) \leqslant \limsup_{n \to \infty} \frac{P(\hat{\mathbb{G}}_n^* \notin S^\delta)}{\eta} \leqslant \varepsilon$$

implying that $P(\hat{\mathbb{G}}_n^* \notin S^\delta | X_1^n) = o_P(1)$. Further, note that

$$\mathbb{E}\left(\left| \hat{g}_n(\hat{\mathbb{G}}_n^*, v, s) - g(\hat{\mathbb{G}}_n^*, v, s) \right| \middle| X_1^n \right)$$
$$\leqslant 2M \cdot P(\hat{\mathbb{G}}_n^* \notin S^\delta | X_1^n) + \sup_{b \in S^\delta} |\hat{g}_n(b, v, s) - g(b, v, s)| \quad (1.33)$$

and, therefore,

$$\begin{aligned}
\sup_{v\in K}|\hat{Q}_{2,n}(v)-\hat{Q}_{3,n}(v)| &\leqslant \sup_{v\in K}\sup_{s\in R_n}\mathbb{E}\left(\left|\hat{g}_n(\hat{\mathbb{G}}_n^*,v,s)-g(\hat{\mathbb{G}}_n^*,v,s)\right| \;\middle|\; X_1^n\right)+o_P(1)\\
&\leqslant \sup_{b\in S^\delta}\sup_{v\in K}\sup_{s\in R_n}|\hat{g}_n(b,v,s)-g(b,v,s)|+o_P(1)\\
&\leqslant 2C_M\sup_{s\in K_n^\delta}\left|\left|\hat{\phi}_n'(s)-\phi_0'(s)\right|\right|+o_P(1)
\end{aligned}$$

where $K_n = S + K + R_{l_n,\lambda_n}$. The latter supremum converges in probability to zero by Assumption 1.5.2.

Third, note that $\sup_{v\in K}|\hat{Q}_{3,n}(v)-\hat{Q}_{4,n}(v)| = o_P(1)$ due to the assumed bootstrap consistency, since $\mathcal{G} = \{g(\cdot;v,s) : v \in K, s \in \mathbb{B}\}$ is a family of bounded Lipschitz functions. Indeed, uniformly in v, s:

$$|g(b;v,s)| \leqslant B_M,$$

$$|g(b_1;v,s)-g(b_2;v,s)| \leqslant C_M\cdot C_\phi\cdot||b_1-b_2||.$$

Therefore, the class of functions $\mathcal{G} = \{g(b;v,s) : v \in K, s \in \mathbb{B}\}$ is a subset of the class of bounded Lipchitz functions with Lipchitz constant $C_M \cdot C_\phi$ and bounded by B_M. Therefore

$$\sup_{v\in K}|\hat{Q}_{2,n}(v)-\hat{Q}_{3,n}(v)| \leqslant \sup_{g\in\mathcal{G}}\left|\mathbb{E}(g(\hat{\mathbb{G}}_n^*)|X_1^n)-\mathbb{E}(g(\mathbb{G}_0))\right| = o_P(1).$$

Fourth, $\sup_{v\in K}|Q_{4,n}(v)-Q(v)| = o(1)$, since $Q_{4,n}$ is a uniformly Lipchitz sequence of functions converging point-wise on a compact set. Indeed, for all n and all $v \in K$, $Q_{4,n}(v)$ is bounded by B_M. Moreover, uniformly in $b, s \in \mathbb{B}$,

$$\begin{aligned}
|g(b,v,s)-g(b,v',s)| &\leqslant C_M\left(||\phi_0'(b+v_1+s)-\phi_0'(b+v_1'+s)||+||v_2-v_2'||\right)\\
&\leqslant C\,||v-v'||,
\end{aligned}$$

and therefore

$$|Q_{4,n}(v)-Q_{4,n}(v')| \leqslant \sup_{b\in\mathbb{B}}\sup_{c\in\mathbb{B}}|g(b,v,s)-g(b,v',s)| \leqslant C\,||v-v'||,$$

so that $\{Q_{4,n}\}$ is a uniformly Lipschitz sequence of functions. For the pointwise convergence, first note that $Q_{4,n}(v) \leqslant Q(v)$ for each $v \in K$. To show the reversed inequality, fix a $v \in K$ an any $\varepsilon > 0$. Then, there is $s_0 \in S(\mathbb{G}_0)$ such that

$$\sup_{s \in S(\mathbb{G}_0)} \mathbb{E}(g(\mathbb{G}_0, v, s)) \leqslant \mathbb{E}(g(\mathbb{G}_0, v, s_0)) + \varepsilon \leqslant \sup_{s \in R_n} \mathbb{E}(g(\mathbb{G}_0, v, s)) + C\varepsilon$$

for large enough n and some constant C independent of n, where the second inequality follows from the Lipschitz-continuity of $s \mapsto \mathbb{E}(g(\mathbb{G}_0, v, s))$ (Lemma 1.5) and Assumption 1.5.3. By Lemma 1.4, a uniformly Lipschitz sequence of functions converging pointwise on a compact set also converges uniformly. Therefore, $\sup_{v \in K} |Q_{4,n}(v) - Q(v)| = o(1)$.

It follows from the preceding discussion that $\sup_{v \in K} |\hat{Q}_n(v) - Q(v)| = o_P(1)$ and, therefore, Lemma 1.7 implies that $\vec{d}_H(\hat{V}_n, \mathcal{V}_0) = o_P(1)$, where $\hat{V}_n$ and $\mathcal{V}_0$ denote the sets of minimizers of $\hat{Q}_n$ and $Q(v)$ correspondingly within K.[26] By contiguity, it also holds that $\vec{d}_H(\hat{V}_n, \mathcal{V}_0) = o_{P_{n,h}}(1)$ for any $h \in T(P)$. Now, let $\hat{v}_n \in \hat{V}_n$ be an arbitrary minimizer of $\hat{Q}_n$. By Prohorov's theorem, for any subsequence, there is a further subsequence, denoted by n', such that $\hat{v}_{n'}$ converges weakly under $P_{n',h}$ to some v. Such v satisfies $P(v \in \mathcal{V}_0) = 1$, its distribution does not depend on h, and it is independent from $\mathbb{G}_0$ since any remaining randomness is due to selecting the minimizer, which is done independently of the data. Then, arguing as in the proof of Theorem 1,

$$\liminf_{n' \to \infty} \sup_{h \in I} \mathbb{E}_{P_{n',h}} \left\{ l_M \left(\sqrt{n'}(\hat{\phi}_{n'} - \phi(\theta_{n'}(h))) \right) \right\}$$
$$\leqslant \sup_{h \in T(P)} \mathbb{E} \left\{ l_M \left(\phi_0'(\mathbb{G}_0 + v_1 + \theta_0'(h)) - \phi_0'(\theta_0'(h)) + v_2 \right) \right\}. \quad (1.34)$$

The supremum in the second line of the above display can be equivalently taken over $s \in \theta_0'(T(P))$ and further over the closure of this set in $\mathbb{B}$ (by Lemma 1.3), which is equal to $S(\mathbb{G}_0)$. Let $\mathcal{V}_0$ be the set of minimizers, within a compact set $K \subseteq \mathbb{D} \times \mathbb{B}$, of

$$Q(v) = \sup_{s \in S(\mathbb{G}_0)} \mathbb{E} \left\{ l_M \left(\phi_0'(\mathbb{G}_0 + v_1 + \theta_0'(h)) - \phi_0'(\theta_0'(h)) + v_2 \right) \right\}$$

[26] Here $\vec{d}_H(A, B) = \sup_{a \in A} d(a, B)$ denotes the directed Hausdorff distance.

Assuming that $v_1, v_2 \in \mathcal{V}_0$ and are independent of $\mathbb{G}_0$, it follows from (1.34) that

$$\begin{aligned}
&\liminf_{n'\to\infty} \sup_{h\in I} \mathbb{E}_{P_{n',h}} \left\{ l_M \left(\sqrt{n'}(\hat{\phi}_{n'} - \phi(\theta_{n'}(h))) \right) \right\} \\
&\qquad \leqslant \mathbb{E}_v \left(\sup_{s\in S(\mathbb{G}_0)} \mathbb{E}_{\mathbb{G}_0} \left\{ l_M \left(\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2 \right) \right\} \right) \\
&\qquad\qquad = \inf_{(v_1,v_2)\in K} \sup_{s\in S(\mathbb{G}_0)} \mathbb{E} \left\{ l_M \left(\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2 \right) \right\}. \quad (1.35)
\end{aligned}$$

Since (1.35) holds along subsequences $\hat{v}_{n'}$, it must hold for the entire sequence as well, and by taking the supremum over all finite $I \subseteq T(P)$, the proof is complete.

CHAPTER 2

Selecting Inequalities for Sharp Identification in Models with Set-Valued Predictions

2.1 Introduction

Partially-identified models have dramatically gained popularity in the recent literature. They allow to relax some of the less credible assumptions required for point-identification while delivering sufficiently informative results and revealing a clear connection between the assumptions and conclusions. Naturally, these advantages come with some conceptual and practical challenges. One of them is that the identified set is often described by a very large number of moment inequalities. While excluding inequalities from the analysis may lead to losing identifying information, adding uninformative and/or poorly estimated inequalities may distort inference. Therefore, some inequality selection procedures are required.

In this paper, I propose a criterion for inequality selection in a large class of partially-identified models with set-valued predictions. Specifically, I consider models with the following structure. There is an observed outcome variable $Y \in \mathcal{Y}$, covariates $X \in \mathcal{X}$, latent variables $U \in \mathcal{U}$, and unknown parameters θ. Given X and θ, the model delivers a set of predictions $G(U, X; \theta) \subseteq \mathcal{Y}$. The researcher does not observe $G(U, X; \theta)$ but maintains the assumption that $Y \in G(U, X; \theta)$.[1] Examples of such models include entry games with

[1] Chesher and Rosen (2017) established an equivalent representation: given X, Y, and θ, the model delivers a set of latent variables $G(Y, X; \theta) \subseteq \mathcal{U}$ such that, by assumption, $U \in G(Y, X; \theta)$. In applications, one characterization might be more convenient than the other. The results of this paper apply similarly in both cases.

multiple equilibria (Tamer, 2003; Ciliberto and Tamer, 2009), network formation models (De Paula et al., 2018; Sheng, 2020; Gualdani, 2021), auctions (e.g., Haile and Tamer, 2003; Aradillas-López et al., 2013b), generalized IV models (Chesher and Rosen, 2017, and others), and discrete choice models with heterogeneous or counterfactual choice sets (Manski, 2007; Barseghyan et al., 2021).

Identified sets in such models can be characterized using a special kind of moment inequalities, obtained as follows. Fix some value $X = x$ and consider an arbitrary measurable subset $A \subseteq \mathcal{Y}$. Since $Y \in G(U, x; \theta)$, the event $G(U, x; \theta) \subseteq A$ implies $Y \in A$. Therefore,

$$P(Y \in A | X = x) \geqslant P(G(U, x; \theta) \subseteq A | X = x; \theta), \tag{2.1}$$

for all measurable subsets $A \subseteq \mathcal{Y}$. Then, a parameter value θ is included in the identified set if it satisfies all of the above inequalities. Using the result from Artstein (1983) on distributions of random sets, one can show that the above inequalities exhaust all of the information contained in the data and maintained assumptions, so that the resulting identified set is sharp. In practice, the total number of the inequalities in (2.1) may be very large or even infinite, in which case checking all of them is infeasible and the researcher faces the problem of inequality selection. Additionally, many of these inequalities do not actually add any information, and, to improve performance of the inference procedures, it may be desirable to exclude them from the analysis.[2]

To address inequality selection, following the literature, I focus on core-determining classes (Galichon and Henry, 2011; Chesher and Rosen, 2017; Luo and Wang, 2018; Molchanov and Molinari, 2018). A class of $\mathcal{C}$ of subsets of $\mathcal{Y}$ is core-determining, if verifying the inequalities in (2.1) for all $C \in \mathcal{C}$ is sufficient to conclude that they hold for all $A \subseteq \mathcal{Y}$. I provide a simple analytical criterion to determine if an inequality associated with a certain subset

[2]In general, one can construct examples where two moment inequalities imply the third one in the population, and yet it might desirable to use all three of them in finite samples. For instance, let Y_1, Y_2 denote two random variables such that $\mathbb{V}ar(Y_1) \gg \mathbb{V}ar(Y_2)$, and suppose it is known that $\mathbb{E}(Y_1 - \theta) \leqslant 0$, $\mathbb{E}(Y_2) \leqslant \mathbb{E}(Y_1)$, and $\mathbb{E}(Y_2 - \theta) \leqslant 0$. Although the first two inequalities imply the third one, they involve Y_1 which may be estimated poorly in small samples, so the third inequality can prove valuable for inference.

$A \subseteq \mathcal{Y}$ is redundant, given the other inequalities. In a series of examples, I show that the proposed criterion can substantially reduce the size of the core-determining classes compared with existing literature and, in settings where the total number of inequalities is finite, find the smallest possible core-determining class. I find that the smallest core-determining class depends only on the structure of the correspondence $G(u, x; \theta)$, and does not directly depend on the probability distribution of $u|X = x$ or the value of θ. Therefore, in applications, this class only has to be computed once before carrying out the rest of the analysis. For that purpose, I propose a simple computational procedure using graph propagation techniques.

This paper contributes to the literature on partial identification using random set theory (see Molinari, 2020; Chesher and Rosen, 2020, for a deatiled review). The most closely related papers are Chesher and Rosen (2017), and Luo and Wang (2018). Chesher and Rosen (2017) propose a general analytical criterion for constructing core-determining classes. Their procedure consists of two steps. First, restrict attention to the set of all unions of elements of the support of $G(y,x;\theta)$, and, second, exclude all unions of suitably disjoint "small" sets. In this paper, I show that one can also exclude all intersections of suitably overlapping "large" sets. In a series of examples, I show that this extra step allows to substantially reduce the size of the core-determining class. Moreover, in settings where the outcome space is finite, I demonstrate that the remaining sets form the smallest possible core-determining class. A result similar to the latter also appears in a working paper by Luo and Wang (2018). While a careful comparison suggests that both approaches lead to the same core-determining class, the characterization in this paper is simpler and more intuitive, and allows to compute the smallest core-determining class more efficiently.

The rest of the paper is organized as follows. Section 2.2 presents motivating examples, provides a formal setup and necessary background and recites known results. Section 2.3 presents new theoretical results, provides an algorithm that can be used to efficiently compute the smallest core-determining class, and discusses using redundant inequalities for inference procedures. Section 1.9 concludes.

2.2 Models with Set-Valued Predictions

2.2.1 Motivating Examples

First, I present several examples, which I revisit throughout the paper to fix ideas and illustrate the main results. The first example is a version of a static entry game studied in Tamer (2003), Ciliberto and Tamer (2009), and Beresteanu et al. (2011), among others.

Example 2.1 (Static Entry Game)**.** Consider a static market entry game with N firms. Firm j chooses $Y_j \in \{0,1\}$, where $Y_j = 1$ represents entry, and receives the payoff

$$\pi_j(Y, \varepsilon_j; \theta) = Y_j(\alpha_j + \delta_j(N(Y) - 1) + \varepsilon_j),$$

where $Y = (Y_1, \ldots, Y_N) \in \{0,1\}^N = \mathcal{Y}$ represents entry decisions, $N(Y)$ is the total number of entrants, $U = (\varepsilon_1, \ldots, \varepsilon_N) \in \mathbb{R}^N$, distributed according to some unknown CDF F, are latent payoff shifters, and (α_j, δ_j) are payoff parameters. Covariates can be included in the model by letting $\alpha_j = \alpha(x_j)$, etc., but are omitted here for simplicity. The firms are assumed to have complete information and play a Nash Equilibrium. The researcher observes Y but not U, and does not specify any particular equilibrium selection mechanism in situations where multiple equilibria are possible. Let $\theta = (\alpha, \delta, F)$ denote all of the model's parameters. The model delivers a set-valued prediction corresponding to the set of pure strategy Nash Equilibria:

$$G(U; \theta) = \{Y \in \{0,1\}^N : Y_j = \mathbf{1}(\alpha_j + \delta_j(N(Y) - 1) + \varepsilon_j \geqslant 0), \text{ for all } j = 1, \ldots, N\}.$$

Figure 2.1 presents possible realizations of $G(U; \theta)$ when $N = 2$ and $\delta_j < 0$. The sharp identified set is given by:

$$\Theta_I = \{\theta : P(Y \in A) \geqslant P(G(U; \theta) \subseteq A | \theta) \text{ for all } A \subseteq \mathcal{Y}\},$$

With N players, the cardinality of the outcome space is 2^N, so there are $2^{2^N} - 2$ nontrivial subsets, each corresponding to a moment inequality. As argued in Section 2.3, most of these

inequalities are redundant, but for $N \geqslant 5$, sharp identified sets remain computationally intractable so additional inequality selection is necessary. ■

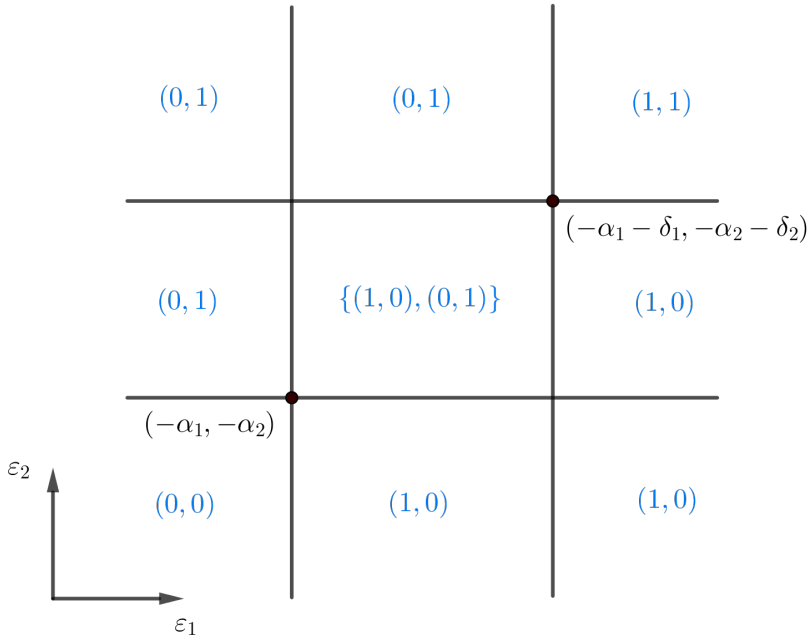


Figure 2.1: Set-Valued Prediction in a Static Entry Model with $N = 2$ and $\delta_j < 0$.

The following example is a version of an English auction model studied in Haile and Tamer (2003), Aradillas-López et al. (2013b), Chesher and Rosen (2017), and Molinari (2020), among others.

Example 2.2 (English Auctions)**.** Consider a symmetric ascending auction with N bidders. For simplicity, assume that there is no reserve price and no minimal bid increment. Let $V_j \in [0, \overline{v}]$ and $B_j \in [0, \overline{v}]$ denote the valuation and bid of player j, and $V_{j:N}$ and $B_{j:N}$ denote the j-th smallest valuation and bid correspondingly. Let F denote the joint distribution of $B = (V_{1:N}, \ldots, V_{N:N})$, which is supported on $S = \{v \in [0, \overline{v}]^N : v_1 \leqslant \cdots \leqslant v_N\}$. The

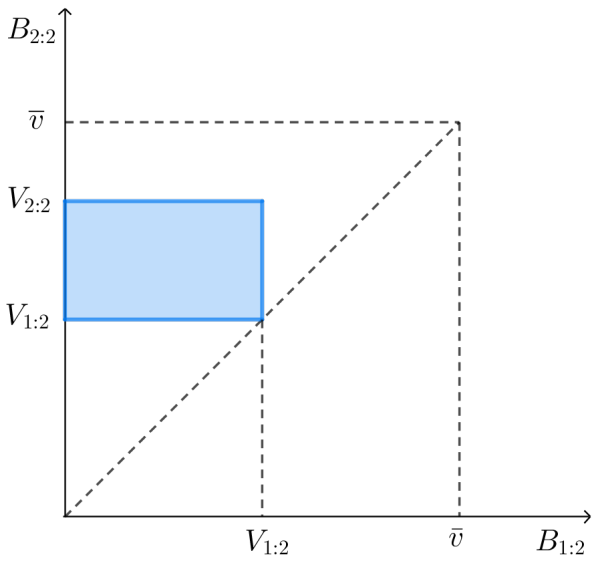


Figure 2.2: Set-Valued Prediction in an English Auction with Two Players

researcher observes $B = (B_{1:N}, \ldots, B_{N:N})$ and wants to learn about $\theta = F$.[3] It is assumed that bidders (i) do not not bid above their valuation, and (ii) do not let their opponents win at an acceptable price. Then, (i) implies $B_{j:N} \leqslant V_{j:N}$ for all j, and (ii) implies $V_{N-1:N} \leqslant B_{N:N}$, so that the set-valued prediction is given by

$$G(V;\theta) = S \cap \prod_{j=1}^{N-1} [0, V_{j:N}] \times [V_{N-1:N}, V_N].$$

Figure 2.2 presents an example realization of $G(V;\theta)$ with $N = 2$. The sharp identified set is given by:

$$\Theta_I = \{\theta : P(B_{1:N} \in A) \geqslant P(G(V;\theta) \subseteq A|\theta) \text{ for all } A \subseteq S\},$$

In this example, the total number of moment inequalities is infinite. As discussed in Section 2.3, the novel core-determining class excludes (infinitely) many redundant inequalities, but additional arguments are required to select among the remaining ones. ■

[3]In symmetric auctions, ordered statistics of bids contain the same amount of information ad the bids themselves. For simplicity, I keep $\theta = F$ as the primitive parameter, even though in practice one is typically interested in simpler objects, such as the marginal distribution of valuations in the IPV setting, or marginal distributions of two highest valuations in the setting with affiliated private values. See Haile and Tamer (2003) and Aradillas-López et al. (2013b) and references therein.

The next example is a discrete choice model with endogeneity and instrumental variables studied in Chesher et al. (2013) and Tebaldi et al. (2019), among others.

Example 2.3 (Discrete Choice with Endogeneity and IV)**.** Consider a model in which individuals choose one of $M+1$ alternatives, $Y \in \{y_0, y_1, \ldots, y_M\}$, and choosing y_j delivers utility $v_j(X) + \varepsilon_j$, where $v_0(X) = 0$, $X \in \mathcal{X}$ are explanatory variables and ε_j are option-specific utility shifters. Individuals are assumed to know the utility of each alternative and choose the one that delivers maximum utility. Some components of X may be correlated with the latent payoff shifters $\varepsilon = (\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_M)$ but the nature of this dependence is left unspecified. The econometrician observes the choice made, $Y = y_{j^*}$, where $j^* = \text{argmax}_j\{v_j(X) + \varepsilon_j\}$, and the explanatory variables X, and has access to instrumental variables $Z \in \mathcal{Z}$ which are independent of ε. Components of Z may either correspond to exogenous components of X or be excluded from the choice problem. In this example, it is more convenient to work in the space of latent variables. Note that $Y = y_j$ if and only if $v_j(X) + \varepsilon_j \geqslant v_k(X) + \varepsilon_k$ for all $k \neq j$. Letting $U_j \equiv \varepsilon_j - \varepsilon_0$ for all $i \neq j$ and $U = (U_1, \ldots, U_M) \in \mathbb{R}^M$ with distribution F, we have that $Y = y_j$ happens if and only if U belongs to the set of latent variables induced by the model:

$$G(y_j, X; \theta) = \begin{cases} \{u : u_j - u_k \geqslant v_k(X) - v_j(X), \text{ for all } k \neq j\} & j \geqslant 1 \\ \{u : u_k < -v_k(X), \text{ for all } k \geqslant 1\} & j = 0 \end{cases}$$

where $\theta = (v_1, \ldots, v_M, F)$. Now, define $G(Y, X; \theta) = \sum_j G(y_j, X; \theta)\mathbf{1}(Y = y_j)$. Since $U \in G(Y, X; \theta)$ and Z is independent from U, we have

$$P(U \in S | Z = z; \theta) \geqslant P(G(Y, X; \theta) \subseteq S | Z = z)$$

for all $z \in \mathcal{Z}$. Figure 2.3 illustrates possible values $G(Y, X; \theta)$ for a given value of X when $M = 2$. The sharp identified set is given by

$$\Theta_I = \{\theta : P(U \in S | Z = z; \theta) \geqslant P(G(Y, X; \theta) \subseteq S | Z = z) \textit{ for all } S \subseteq \mathbb{R}^{M-1}, z \in \mathcal{Z}\}.$$

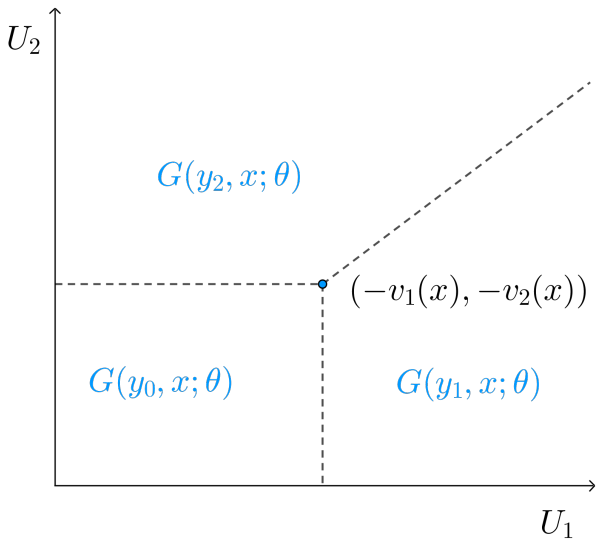


Figure 2.3: Sets of Latent Variables in Discrete Choice Model with $M = 2$.

If X and Z are discrete, the total number of moment inequalities for each θ is finite, but often very large. Identifying redundant inequalities analytically becomes cumbersome, but, if the support of X and Z is of moderate size, the algorithm proposed in Section 2.3 comes to the rescue. If X or Z are continuous, one can bin them or impose further restrictions on $v_k(x)$, such as linearity, to reduce the number of inequalities. In both cases, the resulting identified sets become valid outer regions for Θ_I.

The next example, studied in Manski and Sims (1994); Manski (2003) and Molinari (2020), concerns identifying conditional distributions of interval-observed outcome data.

Example 2.4 (Interval-Observed Outcome Data)**.** Let $Y \in \mathcal{Y} \subseteq \mathbb{R}$ denote the outcome variable and X denote explanatory variables. Suppose that Y is not directly observable, but one observes Y_L and Y_U such that $P(Y_L \leqslant Y \leqslant Y_U) = 1$, and the parameter of interest is the conditional distribution of Y given $X = x$, denoted $\theta = P_{Y|X=x}$. In this example, it is again more convenient to work in the latent-variable space. The set of latent variables induced by the model is $G(Y_L, Y_U) = \mathcal{Y} \cap [Y_L, Y_U]$. Then, for any $A \subset \mathbb{R}$,

$$P(Y \in A|X = x) \geqslant P(G(Y_L, Y_U) \subseteq A|X = x).$$

Letting $\mathcal{P}_{\mathcal{Y}}$ denote the set of all distributions supported on $\mathcal{Y}$, the sharp identified set for θ is:

$$\Theta_I = \{\theta \in \mathcal{P}_{\mathcal{Y}} : \theta(A) \geqslant P(G(Y_L, Y_U) \subseteq A | X = x)\}.$$

As argued below, if $\mathcal{Y} = [\underline{y}, \overline{y}]$, it suffices to consider all intervals $A = [a, b]$ with $\underline{y} \leqslant a \leqslant b \leqslant \overline{y}$. If, additionally, $P(Y_U - Y_L > \kappa | X = x) = 0$, it suffices to consider only "half-lines" $[\underline{y}, a]$, $[b, \overline{y}]$, and "short" intervals $[a, b]$ with $b - a \leqslant \kappa$. Notably, one cannot obtain the sharp identified set using only the marginal distributions of Y_U and Y_L given $X = x$.

2.2.2 Random Sets and Core-Determining Classes

In all of the above examples, the imposed assumptions produce a *random set* of predictions. Naturally, to study such models, it is convenient to employ tools from the theory of random sets. I briefly introduce the necessary concepts below and refer the reader to Molchanov and Molinari (2018) for an accessible textbook treatment. To simplify notation, I abstract away or condition on) the covariates.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\mathcal{Y}, \mathcal{A})$ be a measurable space. A closed random set is a measurable correspondence $G : \Omega \rightrightarrows \mathcal{Y}$ such that each $G(\omega)$ is closed.[4] The support of G, denoted $\text{supp}(G)$, is the set of all possible values of G, i.e., a set of sets. For each $A \in \mathcal{A}$, denote:

$$\begin{aligned} G^-(A) &= \{\omega \in \Omega : G(\omega) \subseteq A\}, \\ G^{-1}(A) &= \{\omega \in \Omega : G(\omega) \cap A \neq \varnothing\}. \end{aligned} \tag{2.2}$$

Then, the distribution of G can be described by its *containment functional* C_G or *capacity functional* T_G, defined as

$$\begin{aligned} C_G(A) &\equiv P(G^-(A)) = P(G \subseteq A), \\ T_G(A) &\equiv P(G^{-1}(A)) = P(G \cap A \neq \varnothing). \end{aligned}$$

[4] Measurability requires $G^{-1}(A) \in \mathcal{F}$ for every closed set $A \in \mathcal{A}$, with $G^{-1}(A)$ defined in (2.2).

Note that $C_G(A) = 1 - T_G(A^c)$. A *selection* of a random set G is any random variable Y that satisfies $Y(\omega) \in G(\omega)$ P-almost surely. The collection of distributions of all selections of G, denoted $\text{Sel}(G)$, is called the *core*.

With these definitions at hand, the identification argument in models with set-valued predictions can be formulated as follows: a parameter value θ is consistent with the data if and only if the observed distribution P_Y of Y belongs to the core of a random set $G(U;\theta)$ for such θ.[5] The following result, due to Artstein (1983), provides a simple characterization.

Lemma 2.1 (Artstein's Inequalities). *Let μ denote a probability distribution on $(\mathcal{Y}, \mathcal{A})$. Then:*

$$\{\mu \in \mathit{Sel}(G)\} \iff \{\mu(A) \geqslant C_G(A), \text{ for all } A \in \mathcal{A}\},$$

or, equivalently,

$$\{\mu \in \mathit{Sel}(G)\} \iff \{\mu(A) \leqslant T_G(A), \text{ for all } A \in \mathcal{A}\}.$$

That is, the core of a random set is completely characterized by moment inequalities with this special structure. In many applications, verifying the inequalities for only a "relatively small" subclass of $\mathcal{A}$ is sufficient. Such subsets are called core-determining classes. The notion originated in Galichon and Henry (2011), and the subsequent contributions include Chesher and Rosen (2017, 2020), and Molinari (2020).

Definition 2.1 (Core-determining Class). *A class $\mathcal{C} \subseteq \mathcal{A}$ of measurable subsets of $\mathcal{Y}$ is core-determining, if for any probability distribution μ on $(\mathcal{Y}, \mathcal{A})$,*

$$\{\mu(A) \geqslant C_G(A) \ \ \forall A \in \mathcal{C}\} \implies \{\mu(A) \geqslant C_G(A) \ \ \forall A \in \mathcal{A}\}.$$

[5] When covariates are present, the requirement is that the conditional distribution $P_{Y|X=x}$ of Y given $X = x$ belongs to the core of a random set $G(U, x; \theta)$ for all $x \in \mathcal{X}$. An equivalent formulation requires that the conditional distribution $P_{U|X=x;\theta}$ of latent variables U given $X = x$ and θ belongs to the core of a random set $G(Y, x; \theta)$ for all $x \in \mathcal{X}$. Appropriate modifications are introduced when instrumental variables Z are present, as in Example 2.3.

In what follows, I propose a new general criterion for finding core-determining classes. The construction proceeds by identifying redundant sets in three consecutive steps, of which the first two are borrowed from Chesher and Rosen (2017). First, for each $A \subseteq \mathcal{Y}$, define a set $\tilde{A} = \bigcup\{C : C \subseteq A, C \in \text{supp}(G)\}$. Then, provided that $\mu(\tilde{A}) \geqslant C_G(\tilde{A})$,

$$\mu(A) \geqslant \mu(\tilde{A}) \geqslant C_G(\tilde{A}) = C_G(A),$$

so that A is redundant given $\tilde{A}$. This means that one can restrict attention to sets that can be written as unions of elements of $\text{supp}(G)$. Second, suppose that, for some $A \in \mathsf{U}_G$, there are sets $A_1, A_2 \subseteq \mathsf{U}_G$ such that $A_1 \cap A_2 = \varnothing$, $A_1 \cup A_2 = A$ and $G^-(A_1 \cup A_2) = G^-(A_1) \cup G^-(A_2)$. Then, provided that $\mu(A_1) \geqslant C_G(A_1)$ and $\mu(A_2) \geqslant C_G(A_2)$,

$$\mu(A) = \mu(A_1) + \mu(A_2) \geqslant C_G(A_1) + C_G(A_2) = C_G(A),$$

so that A is redundant given A_1 and A_2. Combining these two observations yields the following result, which is a version of Theorem 3 in Chesher and Rosen (2017).

Lemma 2.2 (Core-Determining Class from CR17)**.** *Let U_G denote the set of all unions of elements of $\text{supp}(G)$. Let $\mathcal{C} \subseteq \mathsf{U}_G$ be a class of sets such that for every $A \in \mathsf{U}_G \backslash \mathcal{C}$ there exist $A_1, A_2 \in \mathcal{C}$ such that (i) $A_1 \cap A_2 = \varnothing$; (ii) $A_1 \cup A_2 = A$; and (iii) $G^-(A_1 \cup A_2) = G^-(A_1) \cup G^-(A_2)$. Then, $\mathcal{C}$ is core-determining.*

All three conditions of the Lemma are easy to verify in practice. However, since the Lemma places no restrictions on the structure of the random set or underlying probability spaces, the resulting characterization is high-level in a sense that some "hands-on" work is required to see what the resulting inequalities are in practice.

2.3 A New Core-Determining Class

2.3.1 General Case

The core-determining class from Lemma 2.2 can be further refined even further. Suppose that, for some $A \in \mathsf{U}_G$, there are sets $A_1, A_2 \subseteq \mathsf{U}_G$ such that $A_1 \cap A_2 = A$, $A_1 \cup A_2 = \mathcal{Y}$, and $G^-(A) = G^-(A_1) \cap G^-(A_2)$. Then, provided that $\mu(A_1) \geqslant C_G(A_1)$ and $\mu(A_2) \geqslant C_G(A_2)$,

$$1 + \mu(A) = \mu(A_1) + \mu(A_2) \geqslant C_G(A_1) + C_G(A_2) = 1 + C_G(A),$$

implying that $\mu(A) \geqslant C_G(A)$. Therefore, A is redundant given A_1 and A_2. Complementing Lemma 2.2 with this argument yields the first main result of this paper.

Theorem 2.1. *Let U_G denote the set of all unions of elements of $supp(G)$. Let $\mathcal{C} \subseteq \mathsf{U}_G$ denote a class of sets such that, for each $A \in \mathsf{U}_G \backslash \mathcal{C}$ at least one of the following conditions hold:*

1. *There are $A_1, A_2 \in \mathcal{C}$ such that: $A_1 \cap A_2 = \varnothing$, $A_1 \cup A_2 = A$, and $G^-(A_1 \cup A_2) = G^-(A_1) \cup G^-(A_2)$.*
2. *There are $A_1, A_2 \in \mathcal{C}$ such that: $A_1 \cap A_2 = A$, $A_1 \cup A_2 = \mathcal{Y}$, and $G^-(A) = G^-(A_1) \cap G^-(A_2)$, or, equivalently, $G^{-1}(A^c) = G^{-1}(A_1^c) \cup G^{-1}(A_2^c)$.*

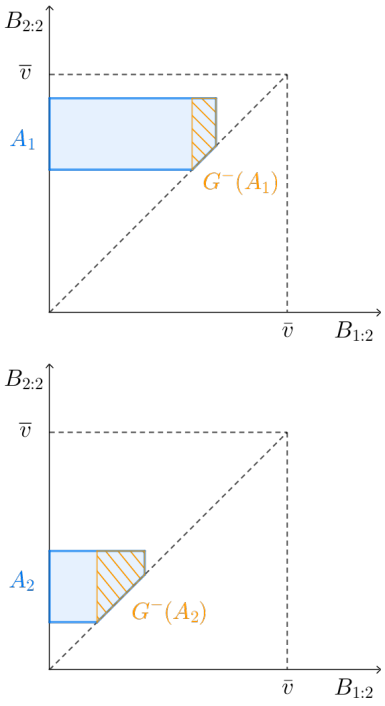
Then $\mathcal{C}$ is core-determining. If there exist a partition $\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2$ such that $\mathcal{Y}_1 \cap \mathcal{Y}_2 = \varnothing$ and $G^{-1}(\mathcal{Y}_1) \cap G^{-1}(\mathcal{Y}_2) = \varnothing$, then, for $A \subseteq \mathcal{Y}_j$ one can take $A_1, A_2 \subseteq \mathcal{Y}_j$.

Heuristically, while Lemma 2.2 states that unions of disjoint "small" sets are redundant, Theorem 2.1 adds that intersections of sufficiently overlapping "large" sets are redundant as well. In many applications, this allows to substantially reduce the size of the core-determining class and may also be beneficial for the inference procedures, as discussed in Section 2.3.3. Below, I illustrate applications of Theorem 2.1 in settings where the total number of inequalities is infinite.

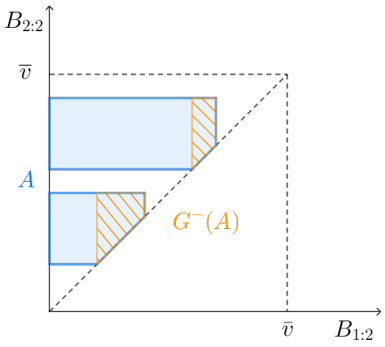
2.3.1.1 Examples Revisited

Example 1 (Continued)**.** For simplicity, consider $N = 2$. Figure 2.4 presents some examples of redundant sets $A \in \mathsf{U}_G$ and sets A_1 and A_2 that satisfying conditions (1) or (2) of Theorem 2.1. In the upper panel, $A_1 \cap A_2 = \varnothing$, $A_1 \cup A_2 = A$, and $G^-(A_1 \cup A_2) = G^-(A_1) \cup G^-(A_2)$. In the lower panel, $A_1 \cap A_2 = A$, $A_1 \cup A_2 = \mathcal{Y}$, and $G^-(A) = G^-(A_1) \cap G^-(A_2)$. These examples generalized to settings with $N > 2$ in a straightforward manner. In this setting, even after deleting all redundant inequalities indentified by Theorem 2.1, one is left with an infinite number of inequalities, and, depending on the specific functional of interest, the resulting identified set may be fairly large. To this end, one can impose additional restrictions such as requiring the distribution of valuations to be positively affiliated, as in Aradillas-López et al. (2013b), independent, as in Haile and Tamer (2003), or explicitly model auction specific heterogeneity as in Chesher and Rosen (2017), among others.

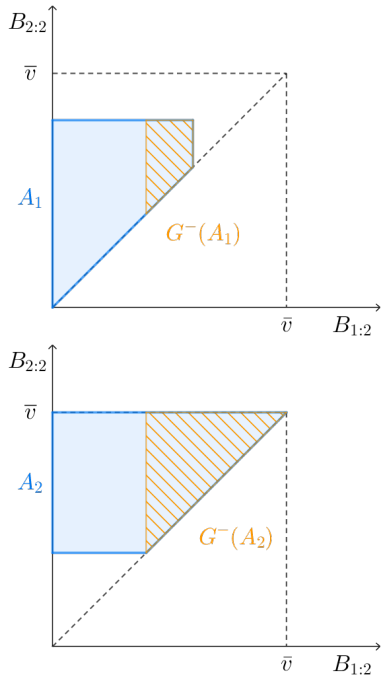
Example 4 (Continued)**.** Assuming that $\mathcal{Y} = [\underline{y}, \overline{y}]$, the set U_G contains all sets that can be expressed as an interval, or a union of disjoint intervals included in $[\underline{y}, \overline{y}]$. Note that for any $A = A_1 \cup A_2 = [a_1, b_1] \cup [a_2, b_2]$ with $a_1 \leqslant b_1 < a_2 \leqslant b_2$, one has $G^-(A_1) \cap G^-(A_2) = \varnothing$. Therefore, such A is redundant given A_1 and A_2. The same argument applies to any other collection of disjoint intervals. Therefore, $\mathcal{C} = \{[a, b] : \underline{y} \leqslant a \leqslant b \leqslant \overline{y}\}$ satisfies Condition 1 of Theorem 2.1, so it is a core-determining class (see Theorem 2.25 in Molchanov and Molinari, 2018). In the absence of other restrictions on the distribution of (Y_L, Y_U), this set cannot be improved. Now, suppose additionally that $P(Y_U - Y_L > \kappa) = 0$ for some κ. Then, any $A = [a, b]$ with $b - a > \kappa$ is redundant, because $A_1 = [\underline{y}, b]$ and $A_2 = [a, \overline{y}]$ satisfy $A_1 \cap A_2 = A$, $A_1 \cup A_2 = \mathcal{Y}$, and $G^{-1}(A^c) = G^{-1}(A_1^c) \cup G^{-1}(A_2^c)$. Therefore, $\mathcal{C} = \{[\underline{y}, a], [a, \overline{y}] : \underline{y} \leqslant a \leqslant \overline{y}\} \cap \{[a, b] : b - a \leqslant \kappa\}$ is core-determining. Note that, in any case, the core cannot be described using only marginal distributions of Y_L and Y_U.



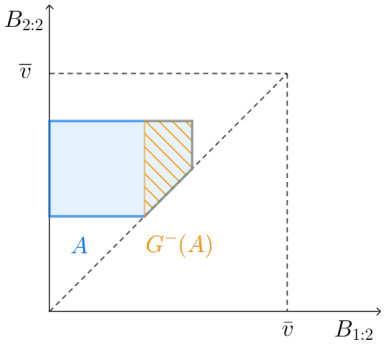
(a) "Small" sets



(b) Redundant union of "small" sets



(c) "Large" sets



(d) Redundant intersection of "large" sets

Figure 2.4: Application of Theorem 2.1 to English Auction Model with Two Players.

2.3.2 Finite Outcome Space

In the settings where the outcome space $\mathcal{Y}$ is finite, the core is characterized by a finite number of inequalities.[6] Then, it is natural to look for the core-determining class with the smallest cardinality. In this section, I show that conditions of Theorem 2.1 actually identify the smallest possible core-determining class, and propose a simple algorithm to compute it in practice.

According to Lemma 2.1, the set of selectionable distributions is a convex polytope defined by a finite number of inequalities:

$$\text{Sel}(G) = \{\mu \in \Delta(\mathcal{Y}) : \mu(A) \geqslant C_G(A) \text{ for all } A \subseteq \mathcal{Y}\},$$

where $\Delta(\mathcal{Y})$ denotes the set of all probability distributions on $\mathcal{Y}$. For $A = \varnothing$ and $A = \mathcal{Y}$, the inequalities hold trivially, so I will exclude them from the analysis below. Identifying non-redundant inequalities is a well-known task in linear programming. Specifically, for every subset $A \subseteq \mathcal{Y}$, define the quantity:

$$\lambda(A) = \min_{p \in \Delta(\mathcal{Y})} \left\{ p(A) \;\middle|\; p(\tilde{A}) \geqslant C_G(\tilde{A}), \text{ for all } \tilde{A} \subseteq \mathcal{Y}, \tilde{A} \neq A \right\}. \tag{2.3}$$

If $\lambda(A) < C_G(A)$, then A cannot be implied by any collection of other inequalities and, therefore, it must belong to any core-determining class. It follows from the literature on redundancy in linear programming (e.g., Telgen, 1983), that the class of all such sets:

$$\mathcal{C}^* = \{A \subseteq \mathcal{Y} : \lambda(A) < C_G(A)\} \tag{2.4}$$

is the smallest core-determining class. Figure 2.5 illustrates.

This characterization illustrates that the smallest core-determining class is well-defined. However, it cannot be used directly for identification arguments. First, even for $\mathcal{Y}$ of relatively small size, solving $2^{|\mathcal{Y}|}$ linear optimization problems may be computationally hard.

[6] Assuming that covariates and instruments are either discrete or conditioned on.

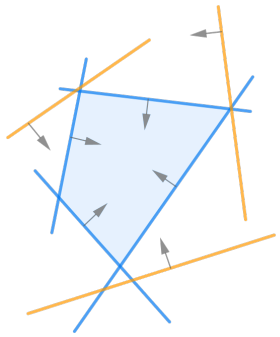


Figure 2.5: Redundant Inequalities Identified by Linear Programming.

Comment: Inequalities depicted in orange are redundant. Inequalities depicted in blue form the smallest core-determining class.

Second, $\mathcal{C}^*$ may depend on θ via $C_G(A) = P(G(U;\theta) \subseteq A)$, meaning that one would have to repeat the procedure for each θ. Below, I provide an alternative characterization that clarifies conditions under which $\mathcal{C}^*$ does not depend on θ, simplifies computation, and directly relates to Theorem 2.1.

Consider a discrete random set $G : \Omega \rightrightarrows \mathcal{Y}$, where $\Omega = \{\omega_1, \ldots, \omega_K\}$ and $\mathcal{Y} = \{y_1, \ldots, y_S\}$, with support $\text{supp}(G) = \{G(\omega_1), \ldots, G(\omega_K)\}$. Such G can be represented with an undirected bipartite graph $\mathbf{B}_G$ with vertices in Ω and $\mathcal{Y}$ and edges (ω_k, y_s) for all $y_s \in G(\omega_k)$. An example is given in Figure 2.6. Conditions of Theorem 2.1 can be translated into the properties of the graph $\mathbf{B}_G$. First, consider sets $A = \{y_1, y_2\}$ and $\tilde{A} = \{y_1\}$. Note that, $A \notin \mathsf{U}_G$, so that A is redundant given $\tilde{A}$ by the argument preceding Lemma 2.2. Note that in this case, the subgraph of $\mathbf{B}_G$ induced by the vertices $(A, G^-(A))$ is disconnected. Second, consider sets $A = \{y_1, y_3\}$, $A_1 = \{y_1\}$, and $A_2 = \{y_3\}$. Note that, $A = A_1 \cup A_2$, $A_1 \cap A_2 = \varnothing$, and $G^-(A_1) \cap G^-(A_2) = \varnothing$. Then, by part (1) of Theorem 2.1, A is redundant given A_2 and A_3. In this case, the subgraph of $\mathbf{B}_G$ induced by $(A, G^-(A))$ is disconnected. Finally, consider sets $A = \{y_3, y_4\}$, $A_1 = \{y_1, y_2, y_3, y_4\}$, and $A_2 = \{y_3, y_4, y_5\}$. Note that $A = A_1 \cap A_2$, $A_1^c = \{y_5\}$, and $A_2^c = \{y_1, y_2\}$, so that $G^{-1}(A_1^c) \cap G^{-1}(A_2^c) = \varnothing$. In this case, the subgraph of $\mathbf{B}_G$ induced by $(A^c, G^{-1}(A^c))$ is disconnected.

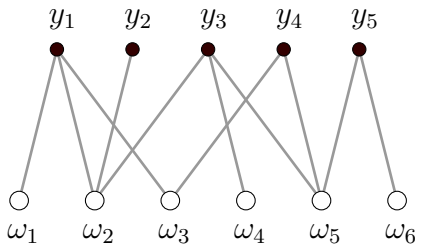


Figure 2.6: Example of a Bipartite Graph Associated with a Random Set.

Theorem 2.1 and the above discussion suggest that all sets A, for which the subgraphs induced by either $(A, G^-(A))$ or $(A^c, G^{-1}(A^c))$ are disconnected, are redundant. The following theorem shows that by eliminating all such sets at once we obtain a core-determining class, which is, in fact, equal to $\mathcal{C}^*$.

Theorem 2.2. *Let $G : \Omega \rightrightarrows \mathcal{Y}$, where $\Omega = \{\omega_1, \ldots, \omega_K\}$ and $\mathcal{Y} = \{y_1, \ldots, y_S\}$ be a discrete random set with a bipartite graph $\mathbf{B}_G$. Suppose that the probability distribution P on Ω satisfies $P(\omega_k) > 0$ for all k. Let $\mathcal{C}$ be a class of subsets $A \subseteq \mathcal{Y}$ such that:*

1. *The subgraph of $\mathbf{B}_G$ induced by $(A, G^-(A))$ is connected;*
2. *The subgraph of $\mathbf{B}_G$ induced by $(A^c, G^{-1}(A^c))$ is connected.*

Then, $\mathcal{C}$ is the smallest core-determining class, i.e., $\mathcal{C} = \mathcal{C}^*$ given by (2.3)–(2.4).

Applications of this Theorem in identification arguments require an auxiliary partitioning of the space of latent variables. Let $P_{U|\theta}$ denote the distribution of latent variables $U \in \mathcal{U}$ for a given θ. If the outcome space is finite, $\mathcal{Y} = \{y_1, \ldots, y_S\}$, so is the support of the random set, $\text{supp}(G(U;\theta)) = \{G_1, \ldots, G_K\}$. Partition the space of latent variables as $\mathcal{U}_\theta = \{u_1, \ldots u_K\}$, where $u_k = \{u \in \mathcal{U} : G(u;\theta) = G_k\}$, and define a measure P_θ on $\mathcal{U}_\theta$ by $P_\theta(u_k) = P_{U|\theta}(\{u : G(u;\theta) = G_k\})$ for all k. Then, instead of $G(U;\theta)$ one can work with a discrete random set $G : (\mathcal{U}_\theta, P_\theta) \rightrightarrows \mathcal{Y}$ defined on a discrete probability space.

Applying Theorem 2.2 with $\Omega = \mathcal{U}_\theta$ and $P = P_\theta$ as defined above, one obtains the smallest set of inequalities exhausting all information about θ. Typically, the support of $G(U;\theta)$, and therefore the network structure of $\mathbf{B}_G$ does not depend on θ, even though the partition $\mathcal{U}_\theta$ and the measure P_θ do. That is, applying Theorem 2.2 for different θ certainly requires calculating the probabilities $P_\theta(u_k)$ and may also require re-labeling of the parts $u_k = \{u : G(u;\theta) = G_k\}$, the smallest set of non-redundant inequalities only has to be computed once.[7] Also note that latent variables $u \in \mathcal{U}$ are typically assumed to have full support, which implies that the induced probability distribution P_θ on $\mathcal{U}_\theta$ will satisfy the assumption of the Theorem.

To compute the class $\mathcal{C}$ from Theorem 2.2 given the bipartite graph $\mathbf{B}_G$, one can use the following algorithm.[8]

Algorithm.

0. If $\mathbf{B}_G$ has several disconnected components, the following steps should be applied to each component separately (still denoting the outcome space by $\mathcal{Y}$).

1. List nontrivial subsets of $\mathcal{Y}$ by their size $k = 1, \ldots, |\mathcal{Y}| - 1$. Initialize $\mathcal{C} = \varnothing$ (or $\mathcal{C} = \mathcal{Y}$ if it is one of the disconnected components).

2. For each $k = 1, \ldots, |\mathcal{Y}| - 1$, do:

   – Pick a subset A with $|A| = k$.

   – If $k \leqslant |\mathcal{Y}|/2$, first check that the subgraph induced by $(A, G^-(A))$ is connected. If so, check that the subgraph induced by $(A^c, G^{-1}(A^c))$ is connected.

   – If $k > |\mathcal{Y}|/2$, do the above in reverse order.

[7] In some cases, as in Example 2.1, no re-labeling is required and the corresponding probabilities have simple closed-form expressions. In other cases, as in Example 2.3, re-labeling and re-calculating probabilities is less straightforward, but can be performed relatively fast.

[8] Calculating the computational complexity of this algorithm and improving upon it is left for further research.

– If both subgraphs are connected, append A to $\mathcal{C}$.

3. Return $\mathcal{C}$.

The following examples illustrate various applications of Theorem 2.2.

2.3.2.1 Examples Revisited

Example 1 (Continued)**.** First, assume that $\delta_j < 0$ so that the firms compete against each other upon entering the market. Then, it is well-known that the set of Nash Equilibria cannot contain two equilibria with different numbers of entrants. The outcome space can be partitioned into disjoint subsets accordingly $\mathcal{Y} = \bigcup_{n=0}^{N} \mathcal{Y}_n$. This property allows to reduce the number of nonredundant inequality dramatically. Specifically, Theorem 2.2 (or Lemma 2.2) immediately implies that all sets of the form $A = \bigcup_{n=0}^{N} A_n$, where $A_n \subseteq \mathcal{Y}_n$, are redundant. Then, the second criterion of Theorem 2.2 can be applied to each subset $\mathcal{Y}_n$ separately, as well as the above Algorithm. In this example, the appropriate partition of the latent variables space is easy to construct; Figure 2.1 illustrates the game with $N = 2$. Note that even though the regions in the partition, as well as their corresponding probabilities, change with θ, the corresponding bipartite graph does not; see Figure 2.7. That is, the core-determining class only needs to be computed once. For $N = 2$, there are 14 inequalities in total and and 5 in the smallest class. For $N = 3$, there are 254 inequalities in total and 15 in the smallest class. For $N = 4$, there are 65534 inequalities in total and 94 in the smallest class. For $N = 5$, the largest $\mathcal{Y}_n$ has approximately 0.6 billion subsets, so the problem is very hard computationally. Next, consider an entry game with complementarities, i.e., $\delta_j > 0$. One can verify that in this case the set of Nash Equilibria only contains equilibria with different numbers of entrants. For $N = 2$ the results are the same as before. For $N = 3$, there are 254 inequalities in total, 85 inequalities in the class from Lemma 2.2, and only 36 inequalities in the smallest class delivered by Theorem 2.2. For $N = 4$, there are 65534 inequalities in total, 18667 inequalities in the class from Lemma 2.2, and only 553 inequalities in the smallest

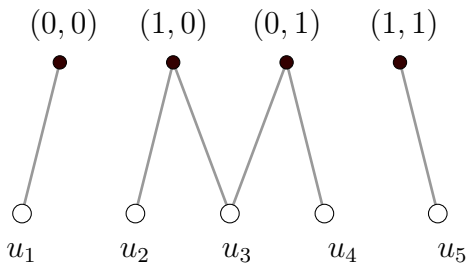


Figure 2.7: Bipartite Graph for the Entry Game in Example 2.1 with $N = 2$

class. $N \geqslant 5$ is computationally infeasible.

Example 2.5 (Network Formation. Adapted from Gualdani (2021))**.** There are N firms forming directed links with each other. For each player j, define a local game Γ_j in which the other $N-1$ players decide whether to link with player j. Let the payoffs in the local game be the same as the payoffs in the entry game with complementarities in Example 2.1. The outcome variable in this example is a network $Y = (Y_{ij})_{i,j=1}^N$ where each $Y_{ij} \in \{0, 1\}$. Then, under appropriate restrictions on payoffs, one can show that a network Y is an Nash Equlibrium (with complete information) of the entire game Γ if and only if $(Y_{ij})_{i \neq j}$ is a Nash Equilibrium in the local game Γ_j, for all j. Assuming additionally that the equilibrium selection rules in local games are independent and letting $\mathcal{C}(\Gamma_j)$ denote a core-determining class in Γ_j, one can show that $\mathcal{C}(\Gamma) = \bigcup_{j=1}^N \mathcal{C}(\Gamma_j)$. To find the smallest core-determining class, one can apply Theorem 2.2 to each Γ_j separately. Then, maintaining the above assumptions, for $N = 3$, there are 254 inequalities in total and 15 in the smallest class. For $N = 4$, there are $\approx 2^{64}$ inequalities in total and only 144 in the smallest class. For $N = 5$, there are $\approx 2^{1024}$ inequalities in total, and only 2212 in the smallest class.

Example 3 (Continued)**.** Consider the case when $y \in \{y_0, y_1, y_2\}$ and $x \in \{x_1, x_2\}$. The first step is to construct an appropriate partition of the space of latent variables. Elements of the partition take the form $\bigcap_{i,j} G(y_i, x_j; \theta)$ where $G(y_j, x_i; \theta)$ is the set of (U_1, U_2) for

which, given x_i, the optimal choice is y_i.[9] Figure 2.8a presents one possible configuration. Numbers $0, 1$, and 2 next to each point indicate the regions of latent variables for which the optimal choice is $y \in \{y_0, y_1, y_2\}$, for a given $x \in \{x_1, x_2\}$. Figure 2.8b presents the corresponding bipartite graph. In this example, the partition may look quite different for different values of $\theta = (v_1(x_1), v_1(x_2), v_2(x_1), v_2(x_2))$, but the structure of the corresponding bipartite graph remains the same up to relabeling of $u_1, \ldots, u_6$. If $x \in \{x_1, \ldots, x_K\}$, there will be $(K+1)(K+2)/2$ elements in the partition, and many different configurations (positions of the points $(v_1(x), v_2(x))$ relative to each other). For example, for $K = 2, 3, 4$, there are 64, 512, and 4096 inequalities in total, while the smallest core-determining class consists only of 12, 42, and 94 inequalities correspondingly.

2.3.3 Redundant Inequalities for Inference

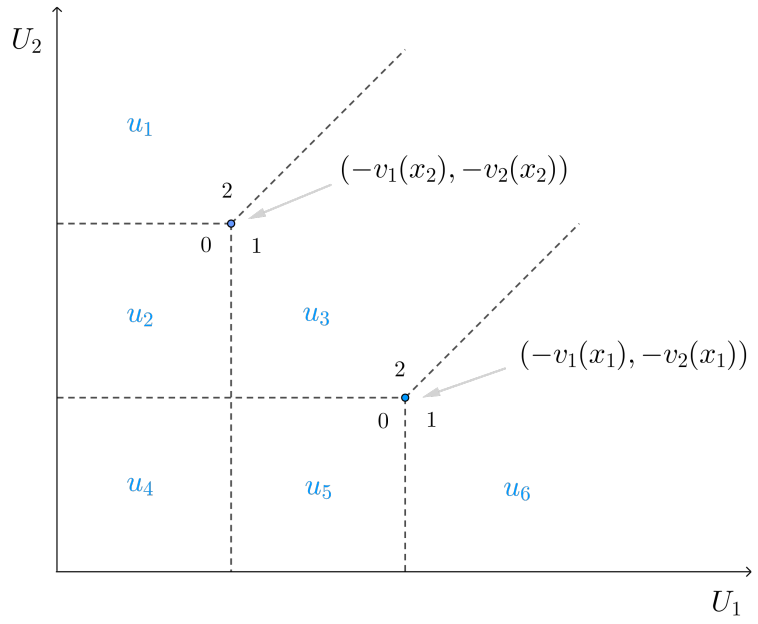
Theorems 2.1 and 2.2 identify inequalities that are redundant for identification. However, for the sake of testing hypotheses and constructing confidence intervals, one does not necessarily want to exclude inequalities deemed redundant in the population if they can be estimated with higher precision than the ones that imply them in finite samples. Here, I investigate whether such inequalities should also be avoided for inference.

The null hypothesis of interest is:

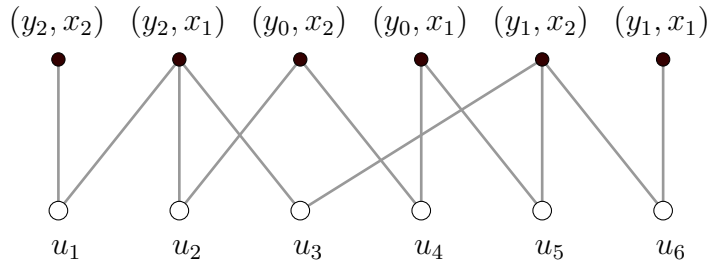
$$H_\theta: \quad P(Y \in A) \geqslant P(U \in G^-(A)|\theta) \text{ for all } A \subseteq \mathcal{Y} \tag{2.5}$$

A variety of different tests for (2.5) have been proposed in the literature, as reviewed in Canay and Shaikh (2017). The tests are based on sample counterparts of the inequalities in (2.5). The left-hand side of each inequality can be estimated by $\hat{P}_n(Y \in A) = n^{-1} \sum_{i=1}^n \mathbf{1}(Y_i \in A)$, while the right-hand side can either be computed analytically or simulated with arbitrary precision by drawing from a hypothesized distribution of latent variables U for a given θ. If

[9] Conceptually, it is the same as the Minimal Relevant Partition from Tebaldi, Torgovitsky, and Yang (2021), although their approach to identification is different.



(a) Partition of The Space of Latent Variables in Example with $x \in \{x_1, x_2\}$



(b) Bipartite Graph with $x = \{x_1, x_2\}$

Figure 2.8: Illustrations for Discrete Choice Model in Example 2.3

using all of the inequalities in (2.5) is practically impossible, suppose that sets $A_0, A_1, \ldots, A_M$ have been selected, and the test will be based on those. In what follows, I denote $p_m = P(Y \in A_m)$, $\hat{p}_{m,n} = \hat{P}(Y \in A_m)$, and $\lambda_m = P(U \in G^-(A_m)|\theta)$ and treat the latter as known.

Consider a test statistic:

$$\hat{T}_n = \max\left\{\max_{0\leqslant m\leqslant M} \frac{\sqrt{n}(\lambda_m - \hat{p}_{m,n})}{\sqrt{\hat{p}_{m,n}(1-\hat{p}_{m,n})}}, 0\right\}. \tag{2.6}$$

The critical values with which to compare $\hat{T}_n$ can be constructed in a number of different ways. To this end, define:

$$J_n(x, s_n) = P\left(\max\left\{\max_{0\leqslant m\leqslant M} \frac{\sqrt{n}(p_m - \hat{p}_{m,n})}{\sqrt{\hat{p}_{m,n}(1-\hat{p}_{m,n})}} + \frac{s_{m,n}}{\sqrt{\hat{p}_{m,n}(1-\hat{p}_{m,n})}}, 0\right\} \leqslant x\right) \tag{2.7}$$

where $s_n = (s_{m,n})_{m=1}^M$. One can derive useful estimators of (2.7) using bootstrap, subsampling, or asymptotic approximation, and the tests proposed in the literature differ in their choice of s_n and the approximation method. Letting $\hat{J}^{-1}_{n,1-\alpha}(s_n)$ denote the corresponding critical value, tests of the form

$$\hat{\phi}_n = \mathbf{1}\left(\hat{T}_n > \hat{J}^{-1}_{n,1-\alpha}(s_n)\right). \tag{2.8}$$

have been shown to achieve uniform size control and can be more or less conservative for specific choices of s_n; see Canay and Shaikh (2017) and Bugni (2016).

Now, suppose that A_0 is redundant given A_1, A_2 in a sense of Theorem 2.1. Should one use such A_0 when testing the hypothesis in (2.5)? Below, I argue that the answer may depend on whether the inequalities corresponding to A_1 and A_2 hold or not. For simplicity, I will omit all other inequalities, and consider only least-favorable tests corresponding to $s_n = 0$ in (2.7). The conclusions remain the same when the remaning inequalities are added, or other valid tests are used.

Let $\hat{\phi}^+_n$ be the test defined by (2.6)–(2.8) using $m = 0, 1, 2$, and $\hat{\phi}^-_n$ be defined similarly using only $m = 1, 2$. It is instructive to compare the two tests is in terms of local asymptotic power functions. Let $\mathbf{P}^\theta_0 = \{P_Y \in \mathbf{P} : H_\theta \text{ holds}\}$ and consider a sequence of distributions

$P_{h,n}$, indexed by a parameter h, approaching some distribution on the boundary of $\mathbf{P}_0^\theta$, i.e., a distribution for which some of the inequalities in (2.5) are binding. The local asymptotic power function is defined as:

$$\kappa(h) = \lim_{n\to\infty} P_{h,n}(\phi_n = 1).$$

Denoting $\sigma_m = \sqrt{\lambda_m(1-\lambda_m)}$ for $m = 0, 1, 2$, consider a sequence $P_{h,n}$ such that $\sqrt{n}(\lambda_m - p_{m,n})/\sigma_m = h_m$, for $m = 1, 2$. Then, $\sqrt{n}(\lambda_{0,n} - p_{0,n})/\sigma_0 = \sigma_1 h_1/\sigma_0 + \sigma_2 h_2/\sigma_0$. Moreover, note that $P_{n,h} \in \mathbf{P}_0^\theta$ for $h_1, h_2 \leqslant 0$, and $P_{n,h} \notin \mathbf{P}_{0,\theta}$ otherwise. Then, an application of Lindeberg-Feller CLT, Slutsky's Theorem, and Continuous Mapping Theorem yields:

$$\begin{aligned} \kappa^+(h) &= P\left(Z^+_{\max}(h) > Q_{1-\alpha}(Z^+_{\max}(0))\right) \\ \kappa^-(h) &= P\left(Z^-_{\max}(h) > Q_{1-\alpha}(Z^-_{\max}(0))\right) \end{aligned},$$

where

$$\begin{aligned} Z^+_{\max}(h) &= \max\left\{Z_1 + h_1, Z_2 + h_2, \tfrac{\sigma_1}{\sigma_0}(Z_1 + h_1) + \tfrac{\sigma_2}{\sigma_0}(Z_2 + h_2), 0\right\} \\ Z^-_{\max}(h) &= \max\left\{Z_1 + h_1, Z_2 + h_2, 0\right\} \end{aligned},$$

$Q_{1-\alpha}(\cdot)$ denotes the $1-\alpha$-quantile of its argument, and (Z_1, Z_2) have joint Normal distribution with $\mathbb{E}(Z_m^2) = 1$ for $m = 1, 2$, and $\mathbb{E}(Z_1 Z_2) = -(\lambda_1\lambda_2)^{1/2}/((1-\lambda_1)(1-\lambda_2))^{1/2}$ for $\lambda_1 + \lambda_2 < 1$ and $\mathbb{E}(Z_1 Z_2) = -((1-\lambda_1)(1-\lambda_2))^{1/2}/(\lambda_1\lambda_2)^{1/2}$ for $\lambda_1 + \lambda_2 > 1$.

Note that, for $h \leqslant 0$, one has $Z^*_{\max}(h) \leqslant Z^*_{\max}(0)$ and therefore $\kappa^*(h) \leqslant \alpha$ for $* \in \{+, -\}$. Moreover, if $h_1, h_2 \ll 0$, it is very likely that $Z^+_{\max}(h) = Z^-_{\max}(h)$, and therefore $\kappa^-(h) > \kappa^+(h)$. Finally, it can be shown that the region of values of h for which $\kappa^-(h) < \kappa^+(h)$ is contained in the half-space $\{h : h_1 + h_2 \geqslant 0\}$. Otherwise, it is hard to draw conclusions analytically, so I refer to simulations.

I compare the power functions using $N = 10^5$ draws of (Z_1, Z_2), for six different values of λ_1, λ_2. The regions of interest are depicted in Figure 2.9 and depend on the relations between σ_0, σ_1, and σ_2. Dashed lines with slopes 2 or 1/2 define the regions where $\sigma_1 \leqslant \sigma_0$ and $\sigma_2 \leqslant \sigma_0$ correspondingly. Note that the blue region corresponds to case (1) of Theorem 2.1, and the orange region corresponds to case (2). Simulation results presented in Figure

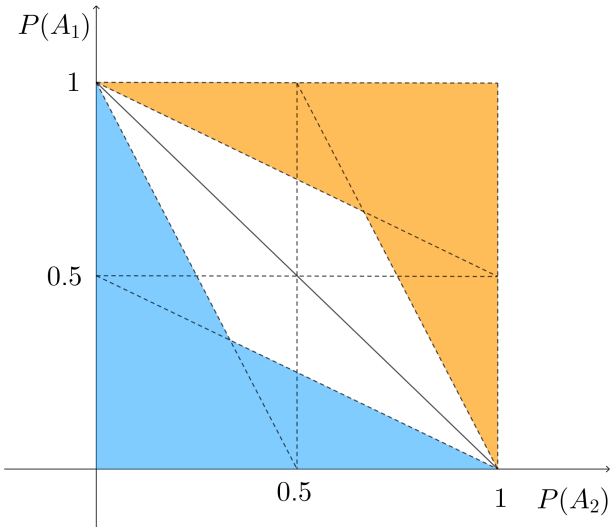


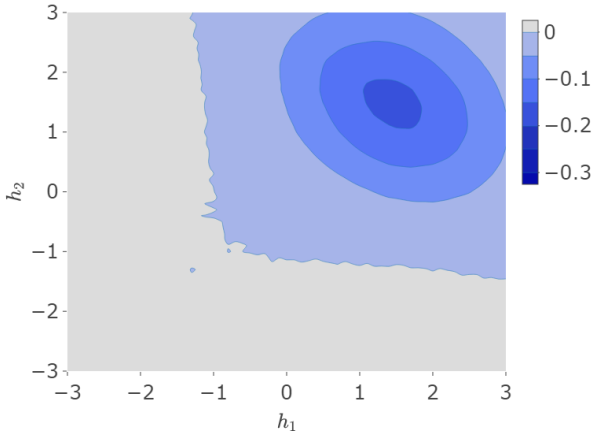
Figure 2.9: Regions of values of (λ_1, λ_2) used in local power comparisons.

2.10, where each plot depicts contour sets for $\kappa^-(h) - \kappa^+(h)$, suggest interesting takeaways that may be used to motivate inequality selection for inference in finite samples.

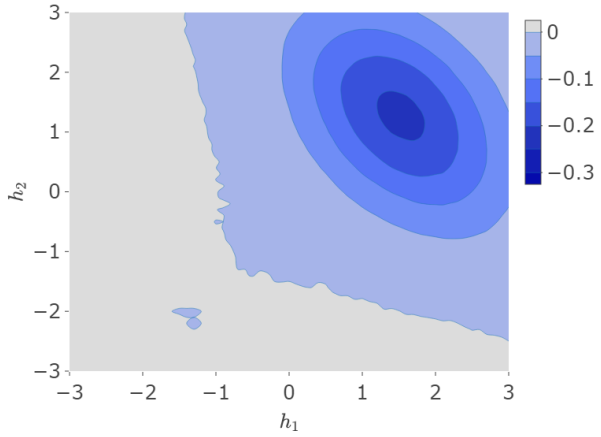
First, note that for $h_1, h_2 \leqslant 0$ and all values of λ_1, λ_2, one has $\alpha \geqslant \kappa^-(h) \geqslant \kappa^+(h)$. This may suggest that adding a redundant inequality when both of the non-redundant inequalities support H_θ can only make the test unnecessarily conservative. On the other hand, for all $h_1, h_2 > 0$ and all values of λ_1, λ_2, one has $\kappa^+(h) > \kappa^-(h)$. This may suggest that adding a redundant inequality when both of the non-redundant inequalities are violated can only increase the power of the test, even if the redundant inequality is estimated relatively imprecisely. Proposing a finite-sample inequality selection procedure that would take the above conclusions into account is an interesting direction for further research.

2.4 Conclusion

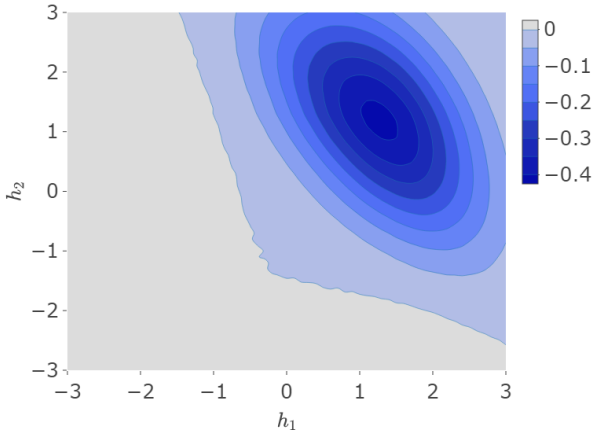
A common practical problem in many partially-identified models is that sharp identified sets are characterized by a very large (or infinite) number of moment inequalities. At the same time, many of those inequalities may be redundant. To motivate and guide inequality



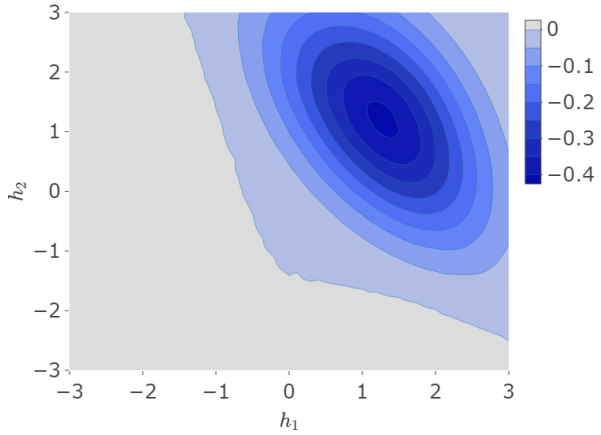
(a) $(\lambda_1, \lambda_2) = (0.2, 0.2)$



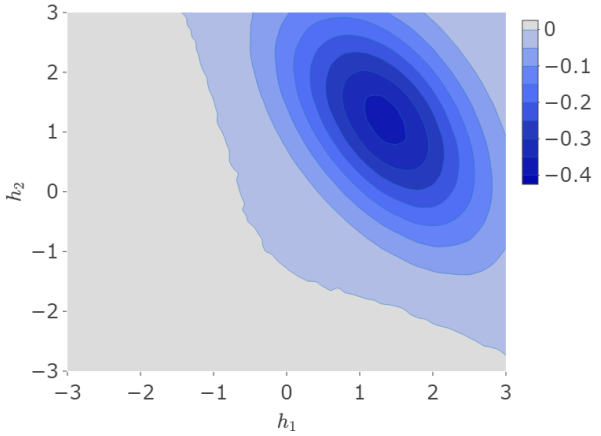
(b) $(\lambda_1, \lambda_2) = (0.1, 0.6)$



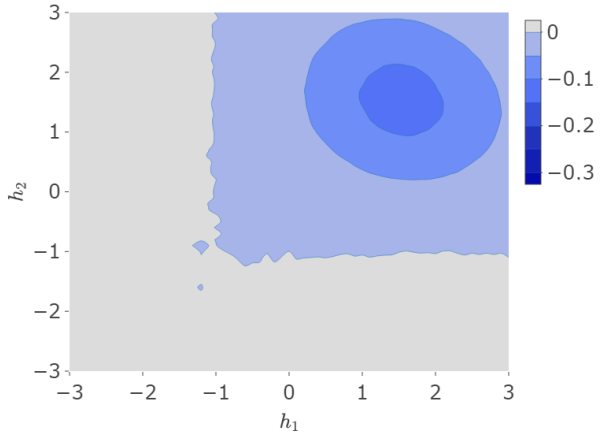
(c) $(\lambda_1, \lambda_2) = (0.4, 0.4)$



(d) $(\lambda_1, \lambda_2) = (0.6, 0.6)$



(e) $(\lambda_1, \lambda_2) = (0.8, 0.4)$



(f) $(\lambda_1, \lambda_2) = (0.9, 0.9)$

Figure 2.10: Local Power Comparisons. Contour sets for $\kappa^-(h) - \kappa^+(h)$.

selection, the literature has focused on characterizing core-determining classes, i.e., sets of inequalities of smaller size that suffice for extracting all of the information from the data and maintained assumptions. In this paper, I proposed a novel general analytical criterion for constructing core-determining classes, and illustrate its utility in a series of popular applications. In settings where the outcome space is finite, I showed that no further improvement is possible and provided an efficient algorithm to compute the smallest possible core-determining classes in applications.

2.5 Appendix: Proofs from the Main Text

2.5.1 Auxiliary Lemmas

Lemma 2.3 (Farkas Lemma)**.** *Let A be an arbitrary $n \times m$ real matrix and b be an arbitrary $n \times 1$ real vector. Then exactly one of the following holds*

1. *There is $\lambda \geqslant 0 \in \mathbb{R}^m$ such that $A\lambda = b$*
2. *There is $\mu \in \mathbb{R}^n$ such that $\mu' A \geqslant 0$ and $\mu' b < 0$*

That is, either a vector b belongs to the convex cone of the columns of A or it can be separated from the cone by a hyperplane. I apply Farkas Lemma to show the following result.

Lemma 2.4. *Let A be an arbitrary $n \times m$ real matrix and b be an arbitrary $n \times 1$ real vector. Let $C = \{1, \ldots, K\}$ with $K < n$, A_C be a submatrix of A composed from the first K rows, and b_C be the corresponding subvector of b. Let a_k' denote the k-th row of A, and $\mathbf{1}_m \in \mathbb{R}^m$ denote a vector of ones. Then, the following are equivalent:*

(1)

$$\{x \in \mathbb{R}^m_+ \mid Ax \geqslant b,\, x'\mathbf{1}_m = 1\} = \{x \in \mathbb{R}^m_+ \mid A_C x \geqslant b_C,\, x'\mathbf{1}_m = 1\}$$

(2) For every $j \notin C$, there is $\lambda_1, \ldots, \lambda_K \geqslant 0$ and $\lambda \in \mathbb{R}$ such that

$$\begin{aligned} a_j &\geqslant \sum_{k=1}^K \lambda_k a_k + \lambda \mathbf{1}_m \\ b_j &= \sum_{k=1}^K \lambda_k b_k + \lambda \end{aligned}$$

Proof. Suppose that (1) holds. Fix some $j \notin C$ and apply Farkas Lemma to:

$$\tilde{A} = \underbrace{\begin{bmatrix} a_1 & a_2 & \ldots & a_K & I_m & \mathbf{1}_m & -\mathbf{1}_m \\ -b_1 & -b_2 & \ldots & -b_K & 0_m & -1 & 1 \end{bmatrix}}_{(m+1)\times(K+m+2)}, \qquad \tilde{b} = \underbrace{\begin{bmatrix} a_j \\ -b_j \end{bmatrix}}_{(m+1)\times 1},$$

where I_m denotes the identity matrix of size m. There are two possible cases.

Case 1. There exists $\lambda \in \mathbb{R}_+^{K+m+2}$ such that $\tilde{A}\lambda = \tilde{b}$, that is,

$$\begin{gathered} \lambda_1 a_1 + \cdots + \lambda_K a_K + \sum_{i=1}^m \lambda_{K+i} e_i + \lambda_{K+m+1} - \lambda_{K+m+2} = a_j \\ \lambda_1 b_1 + \cdots + \lambda_K b_K + \lambda_{K+m+1} - \lambda_{K+m+2} = b_j \end{gathered}$$

where e_i is the i-th column of I_m. Denoting $\lambda = \lambda_{K+m+1} - \lambda_{K+m+2}$, since all λ_{K+i} are non-negative, yields

$$\begin{aligned} a_j &\geqslant \sum_{k=1}^K \lambda_k a_k + \lambda \mathbf{1}_m \\ b_j &= \sum_{k=1}^K \lambda_k b_k + \lambda \end{aligned}$$

Case 2. There exists $\mu \in \mathbb{R}^{m+1}$ such that $\mu' \tilde{A} \geqslant 0$ and $\mu' \tilde{b} < 0$, that is,

$$\begin{gathered} \mu_i \geqslant 0 \ \text{ for } i = 1, \ldots, m \\ \sum_{i=1}^m \mu_i = \mu_{m+1} \\ \sum_{i=1}^m \mu_i a_{ki} \geqslant \mu_{m+1} b_k \ \text{ for } k = 1, \ldots, K \\ \sum_{i=1}^m \mu_i a_{ji} < \mu_{m+1} b_j \end{gathered}$$

Note that $\mu_{m+1} > 0$ and define $x = \mu_{m+1}^{-1} \cdot (\mu_1, \ldots, \mu_m) \in \mathbb{R}_+^m$. Then,

$$\begin{gathered} x' a_k \geqslant b_k \\ x' \mathbf{1}_m = 1 \\ x' a_j < b_j \end{gathered}$$

which contradicts (1). Therefore, this case is impossible, and Case 1 must hold.

Next, suppose that (2) holds. It suffices to show that $RHS \subseteq LHS$ in (1). Pick some $x \in RHS$. Since $x \geqslant 0$ and $x'\mathbf{1}_m = 1$, for any $j \notin C$ we have

$$a_j'x \geqslant \sum_{k=1}^{K} \lambda_k(a_k'x) + \lambda(x'\mathbf{1}_m) \geqslant \sum_{k=1}^{C} \lambda_k b_k + \lambda \geqslant b_j$$

Therefore, $x \in RHS$, and the proof is complete. ■

2.5.2 Proof of Theorem 2.1

By the argument preceding Lemma 2.2, the class U_G is core-determining. So, it remains to show that all $A \in \mathsf{U}_G \backslash \mathcal{C}$ are redundant given $\mathcal{C}$. By definition of $\mathcal{C}$, for any such A, at least one of the conditions of the Theorem must hold. If the first condition holds, A is redundant given $\mathcal{C}$ by the argument immediately Lemma 2.2. If the second condition holds, A is redundant given $\mathcal{C}$ by the argument preceding Theorem 2.1. ■

2.5.3 Proof of Theorem 2.2

The inclusion $\mathcal{C}^* \subseteq \mathcal{C}$ is obvious. Indeed, if $A \notin \mathcal{C}$, it must be redundant by one of the arguments preceding Theorem 2.1. Then, it cannot be that $\lambda(A) < C_G(A)$ as defined in (2.3), and, therefore, $A \notin \mathcal{C}^*$.

Next, I will show $\mathcal{C} \subseteq \mathcal{C}^*$, or, equivalently, $(\mathcal{C}^*)^c \subseteq \mathcal{C}^c$. Denote $M = |\mathcal{Y}|$ and $N = 2^M - 2$. Identify each $A \subseteq \mathcal{Y}$ with a vector $\mathbf{1}_A = (\mathbf{1}(y_m \in A))_{m=1}^M \in \{0,1\}^M$ (excluding $A = \varnothing$ and $A = \mathcal{Y}$). Let $\{1, \ldots, K\}$ enumerate the inequalities in $\mathcal{C}^*$. Since $\mathcal{C}^*$ is core-determining,

$$\left\{ x \in \mathbb{R}_+^M \;\middle|\; \mathbf{1}^T x = 1, \mathbf{1}_{A_k}^T x \geqslant P(G^-(A_k)) \text{ for } k = 1, \ldots, K \right\}$$

$$= \left\{ x \in \mathbb{R}_+^M \;\middle|\; \mathbf{1}^T x = 1, \mathbf{1}_{A_j}^T x \geqslant P(G^-(A_j)) \text{ for } j = 1, \ldots, N \right\},$$

where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^M$. For an arbitrary $A_j \notin \mathcal{C}^*$, I will show that $A_j \notin \mathcal{C}$. If $A_j \notin \mathsf{U}_G$, then certainly $A_j \notin \mathcal{C}$, so for the rest of the proof I will assume that $A_j \in \mathsf{U}_G$. By Lemma

2.4, there exist $\lambda_1, \ldots, \lambda_K \geqslant 0$ and $\lambda \in \mathbb{R}$ such that

$$\begin{aligned} \mathbf{1}_{A_j} &\geqslant \sum_{k=1}^{K} \lambda_k \mathbf{1}_{A_k} + \lambda \mathbf{1} \\ P(G^-(A_j)) &= \sum_{k=1}^{K} \lambda_k P(G^-(A_k)) + \lambda \end{aligned} \tag{2.9}$$

Without loss of generality, assume that $\lambda_k > 0$ for all $k = 1, \ldots, K$ (or drop the sets for which $\lambda_k = 0$).

The first step of the proof is to show that the first line in (2.9) holds with equality. Note that (2.9) implies $\mathbf{1}(G(\omega) \in A_j) \geqslant \sum_{k=1}^{K} \lambda_k \mathbf{1}(G(\omega) \subseteq A_k) + \lambda$. Then, writing $P(G^-(A_j)) = \sum_{\omega \in \Omega} P(\omega) \mathbf{1}(G(\omega) \subseteq A_j)$,

$$\sum_{\omega \in \Omega} P(\omega) \left(\mathbf{1}(G(\omega) \subseteq A_j) - \sum_{k=1}^{K} \lambda_k \mathbf{1}(G(\omega) \subseteq A_k) - \lambda \right) = 0.$$

Since each $P(\omega) > 0$ and the terms in parentheses are non-negative, in fact,

$$\mathbf{1}(G(\omega) \subseteq A_j) = \sum_{k=1}^{K} \lambda_k \mathbf{1}(G(\omega) \subseteq A_k) + \lambda. \tag{2.10}$$

Now, suppose that $\mathbf{1}(y \in A_j) > \sum_{k=1}^{K} \lambda_k \mathbf{1}(y \in A_k) + \lambda$ for some y. Since $A_j \in \mathsf{U}_G$, if $y \in A_j$, there is some ω such that $y \in G(\omega) \subseteq A_j$. Then,

$$\begin{aligned} \mathbf{1}(G(w) \subseteq A_j) &= \textstyle\sum_{k=1}^{K} \lambda_k \mathbf{1}(G(\omega) \subseteq A_k) + \lambda \\ &\leqslant \textstyle\sum_{k=1}^{K} \lambda_k \mathbf{1}(y \in A_k) + \lambda \\ &< \mathbf{1}(y \in A_j), \end{aligned}$$

which states that $1 < 1$, a contradiction. If $y \notin A_j$, the same argument applied to any ω such that $y \in G(\omega)$ leads to $0 < 0$, a contradiction. Therefore,

$$\mathbf{1}(y \in A_j) = \sum_{k=1}^{K} \lambda_k \mathbf{1}(y \in A_k) + \lambda. \tag{2.11}$$

Note that since $\lambda_k > 0$, evaluating the above at any $y \notin A_j$ implies that $\lambda \leqslant 0$.

The second step of the proof starts with the *key observation* that for any ω such that $G(\omega) \subseteq A_j$ and all $y \in G(\omega)$, one has $y \in A_k$ if and only if $G(\omega) \subseteq A_k$, for all $k = 1, \ldots, K$.

Indeed, suppose that, for some k, $y \in G(\omega) \cap A_k$ but $G(\omega) \nsubseteq A_k$. Evaluating (2.10) at ω and (2.11) at y yields

$$\begin{aligned} 1 &= \textstyle\sum_{l:G(\omega)\subseteq A_l} \lambda_l + \lambda \\ 1 &\geqslant \lambda_k + \textstyle\sum_{l:G(\omega)\subseteq A_l} \lambda_l + \lambda, \end{aligned}$$

which cannot hold since $\lambda_k > 0$. Therefore, for all ω such that $G(\omega) \subseteq A_j$, it must be either $G(\omega) \subseteq A_k$ or $G(\omega) \subseteq A_j \backslash A_k$.

It remains to consider two cases.

1. Suppose that $\lambda = 0$. Then, (2.11) and the fact that all $\lambda_k > 0$ imply that $A_j = \bigcup_{k=1}^K A_k$, and also that $A_{k_1} \subseteq A_{k_2}$ cannot hold for $k_1 \neq k_2$. By the key observation, for all ω such that $G(\omega) \subseteq A_j$, either $G(\omega) \subseteq A_k$ or $G(\omega) \subseteq A_j \backslash A_k$ must hold. Since $A_k \neq A_j$ for all k, and the sets A_k cannot be nested, it follows that $A_{k_1} \cap A_{k_2} = \varnothing$ and $G^-(A_{k_1}) \cap G^-(A_{k_2}) = \varnothing$ for all $k_1, k_2 \in \{1, \ldots, K\}$, so that the subgraph induced by $(A_j, G^-(A_j))$ must be disconnected.

2. Suppose that $\lambda < 0$. Then, (2.11) and the fact that $\lambda_k > 0$ imply that each element of A_j must be included in at least one A_k, and $\bigcup_{k=1}^K A_k = \mathcal{Y}$. By the key observation, for any ω such that $G(\omega) \subseteq A_j$, either $G(\omega) \subseteq A_k$ or $G(\omega) \subseteq A_j \backslash A_k$ must hold. If $A_j \backslash A_k \neq \varnothing$ for at least one k, there is no $G(\omega)$ that would connect A_k and $A_j \backslash A_k$. Then, the subgraph induced by $(A_j, G^-(A_j))$ must be disconnected. If $A_j \backslash A_k = \varnothing$ for all k, then $A_j \subseteq A_k$ for all k. Then, by (2.11), $\sum_{k=1}^K \lambda_k + \lambda = 1$, and there cannot be any $y \in \bigcap_{k=1}^K A_k$ but $y \notin A_j$, or, in other words, $A_j = \bigcap_{k=1}^K A_k$. Then, one can re-write (2.11) and (2.10) as:

$$\begin{aligned} \mathbf{1}(y \in A_j^c) &= \textstyle\sum_{k=1}^K \lambda_k \mathbf{1}(y \in A_k^c), \\ \mathbf{1}(G(\omega) \cap A_j^c \neq \varnothing) &= \textstyle\sum_{k=1}^K \lambda_k \mathbf{1}(G(\omega) \cap A_k^c \neq \varnothing). \end{aligned}$$

For any ω such that $G(\omega) \cap A_j^c \neq \varnothing$ and all $y \in G(\omega) \cap A_j^c$, the above implies

$$\sum_{k=1}^K \lambda_k (\mathbf{1}(y \in A_k^c) - \mathbf{1}(G(\omega) \cap A_k^c \neq \varnothing)) = 0.$$

Since all $\lambda_k > 0$, the above implies that for all such ω and y, $G(\omega) \cap A_k^c \neq \varnothing$ holds if and only if $y \in A_k^c$. Since $A_j^c = \bigcup_{k=1}^K A_k^c$, the above holds, specifically, for any ω such that $G(\omega) \cap A_k^c \neq \varnothing$ and all $y \in G(\omega) \cap A_k^c$. Therefore, $G(\omega) \cap A_k^c \neq \varnothing$ happens if and only if $G(\omega) \cap A_j^c \subseteq A_k^c$. In words, if $G(\omega)$ "hits" A_k^c, its restriction on A_j^c must be fully included in A_k^c. Now, define $A_{-1} = \bigcap_{k=1}^K A_k$ and note that $A_1^c \cap A_{-1}^c = \varnothing$. Then, by the previous discussion, there cannot be a ω such that $G(\omega) \cap A_1^c \neq \varnothing$ and $G(\omega) \cap A_{-1}^c \neq \varnothing$, meaning that the subgraph induced by $(A_j^c, G^{-1}(A_j^c))$ must be disconnected, and the proof is complete.

CHAPTER 3

Model Selection for Doubly-Robust Policy Learning

3.1 Introduction

When treatment effects are heterogeneous, an important question is to find out how to best assign treatment to individuals based on their observable characteristics, i.e. find a good policy rule $\pi(x)$ that maps a vector of characteristics x to a treatment. For example, a job training program might only benefit workers of certain education level; some drugs may only work on patients of certain age or with certain medical history; variant advertisement styles lift sales differently depending on customer demographics. In these scenarios, the decision maker might want to look beyond the average treatment effect and search for a good policy rule. Given either experiment or quasi-experiment data, researchers can formulate a statistical decision problem and evaluate policy rules by their expected regret (Manski, 2004; Dehejia, 2005; Stoye, 2009; Bhattacharya and Dupas, 2012; Armstrong and Shen, 2015; Kitagawa and Tetenov, 2018; Athey and Wager, 2021; Mbakop and Tabord-Meehan, 2021).

The problem could be described as follows. A population of agents with observed characteristics $X \in \mathcal{X}$ is to be treated according to a rule $\pi : \mathcal{X} \to \mathcal{D}$ selected from a class of available rules (or interventions) $\pi \in \Pi$. Each treatment rule will result in an outcome $Y \in \mathbb{R}$ (interpreted as utility). Our goal is to learn a treatment rule that maximizes the expected value of Y, denoted $V(\pi)$. For that purpose, we attempt to estimate $V(\pi)$ with $\hat{V}_n(\pi)$ using available data on the outcomes Y_i, treatments T_i, covariates X_i, and optional auxiliary variables Z_i. Specifically, we have access to a collection $W_1^n = \{W_i\}_{i=1}^n$ of i.i.d.

samples $W_i = (Y_i, T_i, X_i, Z_i)$ distributed according to $P \in \mathbf{P}$. We note that the space of observed treatments $T_i \in \mathcal{T}$ may not be the same as the space of interventions $\mathcal{D}$.

The literature has pointed out that in many situations, the set of rules that a policy maker can choose from is constrained by various practical concerns such as budget, fairness, or simplicity. We notice that this constrained set of rules, call it Π, could nevertheless be ambiguous to a practitioner. For example, regulations may dictate that only certain variables could be included in the determination of treatment assignment and a decision tree up to depth four should be employed, but whether to use all of the variables and what exact depth of trees to consider is still up to the practitioner to decide. A better policy π would very likely exist in a larger class Π, but a too complex Π might not work well with the limited amount of data. Just like in many statistical estimation problems, there is a trade-off between bias and variance. Hence, picking a right class Π is a model selection problem for the practitioner.

In this chapter, we focus on the following question: if a practitioner can choose between several different classes of policy rules, denoted Π_k for $k \geqslant 1$, which class should they choose? To answer this question, we need a criterion to compare different data-dependent treatment rules. In line with the literature on statistical treatment rules, we evaluate the performance of treatment rules in terms of their expected regret, $\mathbb{E}[R(\hat{\pi}_n)]$, where regret is defined as

$$R(\pi) = \max_{\pi' \in \Pi^*} V(\pi') - V(\pi).$$

relative to some ideal policy class Π^* that may be infeasible, unknown or arbitrarily set. In the aforementioned example, Π^* could be thought as the largest set of rules allowed under the regulation. Now, to see the trade-off in picking the class, let $\hat{\pi}_{n,k}$ denote the optimal treatment rule chosen from a class Π_k, the regret can be written as:

$$R(\hat{\pi}_{n,k}) = \underbrace{\max_{\pi \in \Pi^*} V(\pi) - \max_{\pi \in \Pi_k} V(\pi)}_{\text{Approximation Error}} + \underbrace{\max_{\pi \in \Pi_k} V(\pi) - V(\hat{\pi}_{n,k})}_{\text{Estimation Error}}.$$

Intuitively, we see that: more complex rules have a better chance of reducing the approximation error, but, for a given sample size, might have larger estimation error.

We adapt and extend two recent methods proposed in Mbakop and Tabord-Meehan (2021) and Athey and Wager (2021). Mbakop and Tabord-Meehan (2021) introduces the penalized welfare maximization (PWM) rule, which itself is an extension to the empirical welfare maximization (EWM) rule proposed in Kitagawa and Tetenov (2018). The PWM rule adds penalization to the EWM rule to achieve model selection. The authors establish a finite-sample upper bound on the expected regret of the PWM rule. The bound converges to zero at $n^{-1/2}$ rate, which is proved to be optimal. A limitation to both EMW and PWM is that when the propensity score is unknown and has to be estimated, these methods would no longer be rate-optimal. Athey and Wager (2021) propose a method that could retain the $n^{-1/2}$ rate even with estimated propensity scores by leveraging doubly robust estimation, but their method does not incorporate model selection. In this chapter, we propose a method that could achieve both.

Following the aforementioned two papers, we propose the following procedure to select the best class. Define the penalized empirical welfare function:

$$Q_{n,k}(\pi) = \hat{V}_n(\pi) - \hat{C}_{n,k}(\pi),$$

where $\hat{V}_n(\pi)$ is a doubly robust estimate of $V(\pi)$ and $\hat{C}_{n,k}(\pi)$ represents a penalty for model complexity, which, informally speaking, estimates how much the model overfits the data. For each k, solve for

$$\hat{\pi}_{n,k} = \underset{\pi \in \Pi_k}{\operatorname{argmax}} \hat{V}_n(\pi),$$

choose

$$\hat{k} = \underset{k}{\operatorname{argmax}} Q_{n,k}(\hat{\pi}_{n,k}),$$

and set

$$\hat{\pi}_n \equiv \hat{\pi}_{n,\hat{k}}.$$

Our main result is to show that such $\hat{\pi}_n$ is adaptive in a sense that it automatically picks up the "right" class and has the optimal rate of convergence in terms of expected regret. Our regret bounds hold in finite samples, are tighter than the bounds available in the literature

and easily generalize to arbitrary discrete policy rules. Moreover, since the welfare estimation $\hat{V}_n(\pi)$ is based on doubly robust scores, our method retains the optimal $n^{-1/2}$ rate in general setups including quasi experiments where the propensity scores have to be estimated.

In Section 3.2, we further describe the setup and introduce our assumptions. In Section 3.3, we revisit known results from the literature and present modified and refined versions of them. Our main results are in Section 3.4, where we formally introduce our new algorithm, the robust penalized welfare maximization (RPWM) rule. We present bound on expect regret of the RPWM rule and prove that it is rate-optimal. Section 3.5 presents a simulation study and Section 3.6 concludes. Proofs are collected in the Appendix, Section 3.7.

3.2 Setup

We consider the standard potential outcomes framework (Neyman, 1923; Rubin, 1974). Specifically, let $Y_i(t)$ denote an outcome that we would have observed if the treatment had been set to $T_i = t$, and $Y = Y(T)$ denote the observed outcome. Let $\theta = \mathbb{E}[\tau(X)]$ denote the average treatment effect. Our main assumption, following Athey and Wager (2021) and Chernozhukov et al. (2016), is that we can identify θ via a doubly-robust moment condition.

Assumption 3.2.1 (Identification)**.** *Let $m(x,t) = \mathbb{E}_P[Y(t)|X = x] \in \mathcal{M}$. Assume that $m(x,t)$ induces a treatment effect function $\tau_m(x,t)$ such that:*

1. *The welfare function can be expressed as $V(\pi) = \mathbb{E}_P[\pi(X)\tau(X)]$, where $\tau(X) = \mathbb{E}_P[\tau_m(X,T)|X]$.*

2. *The map $m \mapsto \tau_m$ is linear and there is a weighting function $g(x,z)$ such that for any $\tilde{m}(x,t) \in \mathcal{M}$*

$$\mathbb{E}_P[\tau_{\tilde{m}}(X,T) - g(X,Z)\tilde{m}(X,T)|X] = 0.$$

The auxiliary variable Z could be an instrumental variable, or equals to X when X is exogeneous. We illustrate this setting with three important examples borrowed from (Athey

and Wager, 2021).

Example 3.1 (Binary Treatments with Selection on Observables)**.** Under conditional ignorability assumption $T \perp (Y(1), Y(0))|X$, condition 2 in Assumption 3.2.1 is satisfied with

$$g(x,t) = \frac{t - e(x)}{e(x)(1-e(x))}, \quad \tau_m(x) = m(x,1) - m(x,0),$$

where $e(x) = P(T = 1|X = x)$ is the propensity score. Then the welfare function is

$$V(\pi) = \mathbb{E}_P[\pi(X)\tau(X)] = \mathbb{E}_P[Y(\pi(X))] - \mathbb{E}_P[Y(0)],$$

which corresponds to our utilitarian welfare objective.

Example 3.2 (Endogenous Binary Treatments with Binary Instruments)**.** Assume that Z is a valid instrument conditional on X in the sense of Assumption 2.1 of Abadie (2003), and further assume that conditional average treatment effect equals conditional local average treatment effect, then we can have

$$\tau_m(x) = m(x,1) - m(x,0) = \frac{\mathbb{C}ov[Y, Z|X = x]}{\mathbb{C}ov[T, Z|X = x]}.$$

Then condition 2 in Assumption 3.2.1 is satisfied with

$$\begin{aligned} g(x,z) &= \frac{1}{\Delta(x)} \frac{z - \Xi(x)}{\Xi(x)(1-\Xi(x))}, \\ \Xi(x) &= P[Z = 1|X = x], \\ \Delta(x) &= P[W = 1|Z = 1, X = x] - P[W = 0|Z = 1, X = x] \end{aligned}$$

Since $\tau_m(x)$ is the same as in the Example 3.1, the resulting welfare function is the same.

Example 3.3 (Continuous Treatments)**.** Suppose the treatment variable T is continuous and exogenous, i.e. $\{Y(t)\} \perp T|X$, then we let

$$\tau_m(x,t) = \frac{d}{dv} m(x, t+v)\Big|_{v=0}.$$

Under regularity conditions, condition 2 of Assumption 3.2.1 is then satisfied with a function $g(X,T)$ derived via integration by parts (Powell et al., 1989)

$$\int\int \frac{d}{dt} m(X,T)\Big|_{t=T} dF_{T|X} dF_X = \int\int \frac{d}{dt} g(X,T) m(X,T) dF_{T|X} dF_X,$$
$$g(X,T) = -\frac{d}{dt}\log(f(t|X))\Big|_{t=T}.$$

In this case the welfare function is

$$V(\pi) = \frac{d}{dv}\mathbb{E}[Y(T + v\pi(X))]\Big|_{v=0},$$

which is the average effect of a nudge following policy $\pi(x)$.

In the above settings, Chernozhukov et al. (2016) proposed estimating θ by

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1} \hat{\Gamma}_i, \quad \hat{\Gamma}_i = \tau_{\hat{m}}(X_i, T_i) + \hat{g}(X_i, Z_i)(Y_i - \hat{m}(X_i, T_i)),$$

where $\hat{g}(\cdot)$ and $\hat{m}(\cdot)$ are preliminary estimates of the nuisance functions $g(\cdot)$ and $m(\cdot)$. Using cross-fitting and Neyman orthogonality of the moment condition $\theta = \mathbb{E}[\Gamma(W; m, g)]$, the authors show that $\hat{\theta}_n$ is $\sqrt{n}$-consistent and asymptotically Normal, provided that $\hat{g}(\cdot)$ and $\hat{m}(\cdot)$ converge sufficiently fast, and may also be semiparametrically efficient (Newey, 1994b).

Athey and Wager (2021) proposed using the orthogonal scores $\hat{\Gamma}_i$ for policy learning. Specifically, under Assumption 3.2.1, $V(\pi) = \mathbb{E}(\pi(X)\Gamma(W))$, so that a feasible sample analog can be constructed as $\hat{V}_n(\pi) = n^{-1}\sum_{i=1}^n \pi(X_i)\hat{\Gamma}_i$. Then, by establishing that $\hat{V}_n(\pi)$ approximates $V(\pi)$ uniformly well over $\pi \in \Pi$, Athey and Wager (2021) show that $\hat{\pi}_n = \text{argmax}_{\pi\in\Pi} \hat{V}_n(\pi)$ is rate-optimal in terms of expected regret.[1]

In section 3.4, we propose our method that complements their results with model selection. Specifically, we propose a procedure that selects the "best" class of treatment rules to choose from in a data-driven fashion. It resolves the trade-off between approximation and

[1] Athey and Wager (2021) work with $A(\pi) = 2V(\pi) - \mathbb{E}(\tau(X)) = \mathbb{E}((2\pi(X) - 1)\tau(X))$ and its feasible analog, but the modification here changes neither the problem nor the solution.

estimation error described earlier and can also be extended to handle policy classes of infinite VC-dimension as in Mbakop and Tabord-Meehan (2021). Another difference between our results to theirs is that our bounds hold in finite sample while they derived asymptotic bounds.

Now, we state the high-level assumptions on the first stage estimation that provides us with $\hat{\Gamma}_i$.

Assumption 3.2.2 (DGP and First-stage Estimators). *In the setting of Assumption 3.2.1, assume that* $\mathbb{E}_P[m^2(X,T)] \vee \mathbb{E}_P[\tau_m^2(X,T)] \vee \mathbb{E}[g^2(X,Z)] < \infty$, *and we have access to estimators* $\hat{m}(x,d)$, $\tau_{\hat{m}}(x,d)$, *and* $\hat{g}(x,z)$ *depending on the data* W_1^n *and satisfying the following conditions. For some* $0 < \zeta_m, \zeta_g < 1$, *with* $\zeta_m + \zeta_g \geqslant 1$, *and a positive sequence* $a(n) \to 0$ *as* $n \to 0$,

$$\mathbb{E}_P[(\hat{m}(X,T) - m(X,T))^2] \vee \mathbb{E}_P[(\tau_{\hat{m}}(X,T) - \tau_m(X,T))^2] \leqslant \frac{a(n)}{n^{\zeta_m}},$$

$$\mathbb{E}_P[(\hat{g}(X,Z) - g(X,Z))^2] \leqslant \frac{a(n)}{n^{\zeta_g}},$$

where (X,D,Z) *is an independent test sample drawn from* P, *for all* $P \in \mathbf{P}$.

The above assumptions on first stage estimation is weaker than the equivalent in Athey and Wager (2021) as we do not assume uniform consistency. Next, we assume that the policy classes have finite VC dimensions.

Assumption 3.2.3 (Policy Rules). *The class of available policy rules is* $\Pi = \bigcup_{k=1}^{K} \Pi_k$, *for some finite* K, *and each* Π_k *has a finite VC dimension denoted* $VC(\Pi_k)$. *The no-treatment rule,* $\pi(x) = 0$ *for all* $x \in \mathcal{X}$, *is included in each* Π_k.

At last, we assume that the function $g(x,z)$ is bounded away from zero.

Assumption 3.2.4 (Overlap Condition). *There is an* $\eta > 0$ *such that the weighting function satisfies* $\sup_{x,z} |g(x,z)| \leqslant \eta^{-1}$ *for all* $P \in \mathbf{P}$.

3.3 Related Results

In this section, we revisit some closely related known results in the literature and present modified and improved version of them.

First, we revisit regret bounds of Kitagawa and Tetenov (2018). Assume that we are in the setting of Example 3.1 and the propensity score is known. Then, the welfare can be expressed as

$$V(\pi) = \mathbb{E}\left[\pi(X)\left(\frac{YT}{e(X)} - \frac{Y(1-T)}{1-e(X)}\right)\right],$$

where $e(X) = P(T = 1|X)$ denotes the propensity score, with a sample counterpart

$$\hat{V}_n^E(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\left(\frac{Y_i T_i}{e(X_i)} - \frac{Y_i(1-T_i)}{1-e(X_i)}\right). \tag{3.1}$$

Kitagawa and Tetenov (2018) consider the Empirical Welfare Maximization (EWM) rule, defined as

$$\hat{\pi}_n^{EWM} = \underset{\pi\in\Pi}{\operatorname{argmax}}\, \hat{V}_n^E(\pi).$$

They derive the upper bound on the worst-case expected regret of this rule over all distributions with bounded outcomes and propensity scores. In the following theorem, we extend and sharpen their result allowing for unbounded outcomes.[2] Define a set of distributions:

$$\mathcal{P}_{B,\eta} = \{P \in \mathbf{P} : \eta \leqslant P(T = 1|X) \leqslant 1 - \eta \text{ a.s.}, \mathbb{E}_P[Y^2] \leqslant B^2\}.$$

Theorem 3.1 (EWM Revisited)**.** *Assume that treatments are binary, $T = \{0, 1\}$, unconfoundedness holds, $(Y(0), Y(1)) \perp T|X$, and the propensity score $e(X)$ is known. Let $\hat{\pi}_n^{EWM} = \operatorname{argmax}_{\pi\in\Pi} \hat{V}_n^E(\pi)$, with $\hat{V}_n^E(\pi)$ defined in (3.1), denote the EWM rule. Then,*

$$\sup_{P\in\mathcal{P}_{B,\eta}} \mathbb{E}_P[R(\hat{\pi}_n^{EWM})] \leqslant C\frac{B}{\eta}\sqrt{\frac{VC(\Pi)}{n}},$$

where $C \leqslant 58$ is a universal constant.

[2] In addition to allowing unbounded outcomes, we obtain a substantially smaller constant. Kitagawa and Tetenov (2018) assume that $Y \in [-M/2, M/2]$ and derive an upper bound of the form $KM\eta^{-1}\sqrt{VC(\Pi)/n}$. A careful examination of the proof of their Theorem 1 suggests that the result holds with $K \approx 68$. To compare, note that for any distribution P such that $Y \in [-M/2, M/2]$, we have $(\mathbb{E}_P[Y^2])^{1/2} \leqslant M/2$. Then, our Theorem 3.1 implies that the expected regret bound holds with $C/2M\eta^{-1}\sqrt{VC(\Pi)/n}$, where $C/2 = 29$.

We complement this result with a tight lower bound to show, in particular, that the EWM rule with known propensity scores is rate-optimal.

Theorem 3.2 (New Regret Lower Bound)**.** *Under the assumptions of Theorem 3.1, for all fixed* $n \geqslant 4VC(\Pi)/\eta$*,*

$$\inf_{\hat{\pi}_n} \sup_{P \in \mathcal{P}_{B,\eta}} \mathbb{E}_P[R(\hat{\pi}_n)] \geqslant 0.07 \cdot \frac{B}{\eta} \sqrt{\frac{VC(\Pi) - 1}{n}} - \kappa_n,$$

where $\kappa_n = 0.14B/\eta \cdot (VC(\Pi) - 1)/n$*, and the right-hand side is positive.*

Remark 3.1 (Unknown Propensity Score)**.** One important limitation of the above result is the assumption that the propensity score is known. If the propensity score is unknown and has to be estimated, one can plug the estimator in (3.1) and maximize the corresponding objective function. Then, a result similar to Theorem 3.1 holds with an additional $O(\phi_n^{-1})$ term, where ϕ_n is the rate of convergence of the propensity score estimator, which is generally slower than $\sqrt{n}$. In such cases, $\hat{\pi}_n^{EWM}$ is no longer rate-optimal.

In the same setting, Mbakop and Tabord-Meehan (2021) propose a treatment rule that accounts for model selection, called Penalized Welfare Maximization (PWM). Here, for simplicity, we only revisit the so-called holdout procedure defined as follows:

1. Let $l = \lceil (1-s)n \rceil$ and $r = n - l$ for some $s \in (0,1)$, and call $W_1, \ldots, W_l$ the estimating sample, and $W_{l+1}, \ldots, W_n$ the test sample. We use subscripts l, r, and n for quantities that depend on the estimating sample only, on the test sample only, and on the entire sample.

2. Compute the EWM rules $\hat{\pi}_{l,k} \equiv \text{argmax}_{\pi \in \Pi_k} \hat{V}_l^E(\pi)$ for each Π_k using the estimating sample. Evaluate each $\hat{\pi}_{l,k}$ by computing the penalized welfare $Q_{n,k}(\hat{\pi}_{l,k}) = \hat{V}_l^E(\hat{\pi}_{l,k}) - \hat{C}_{n,k}$ where the penalty is $\hat{C}_{n,k} = \hat{V}_l^E(\hat{\pi}_{l,k}) - \hat{V}_r^E(\hat{\pi}_{l,k})$.

3. Select $\hat{k} = \text{argmax}_k Q_{n,k}(\hat{\pi}_{l,k})$, and define[3] $\hat{\pi}_n^{PWM} \equiv \hat{\pi}_{n,\hat{k}}$.

[3] A slight abuse of notation here: $\hat{\pi}_{n,\hat{k}}$ is obtained by plugging in $k = \hat{k}$ into $\hat{\pi}_{l,k}$. However, we replace

This procedure is natural: we estimate each $\hat{\pi}_{l,k}$ using the estimating sample, evaluate their performance by computing the empirical welfare on the test sample, $Q_{n,k} = \hat{V}_r(\hat{\pi}_{l,k})$, and select the best estimator. The following result shows that such estimator automatically selects the best class and attains the optimal rate of convergence.[4]

Theorem 3.3 (PWM Revisited)**.** *Assume that treatments are binary, $T = \{0, 1\}$, unconfoundedness holds, $(Y(0), Y(1)) \perp T|X$, and the propensity scores are known. Let $\hat{\pi}_n$ denote the PWM rule computed with the holdout penalty as described above. Then, for any $P \in \mathcal{P}_{B,\eta}$,*

$$\mathbb{E}_P[R(\hat{\pi}_n^{PWM})] \leqslant \inf_{k \leqslant K} \left\{ V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}_P[\hat{C}_{n,k}] \right\} + R_n$$

where V_Π^* and $V_{\Pi_k}^*$ denote the maximum welfare attainable within the corresponding classes (both depend on P), and $R_n = B/\eta \cdot K/\sqrt{sn}$.

Moreover, letting $\mathcal{P}_{B,\eta}^k \subset \mathcal{P}_{B,\eta}$ be a set of distributions such that $V_\Pi^* = V_{\Pi_k}^*$,

$$\sup_{P \in \mathcal{P}_{B,\eta}^k} \mathbb{E}_P[R(\hat{\pi}_n^{PWM})] \leqslant \frac{B}{\eta} \left(C \sqrt{\frac{VC(\Pi_k)}{(1-s)n}} + K \sqrt{\frac{1}{sn}} \right)$$

where $C \leqslant 58$ is a universal constant.

To gain interpretation, recall that selecting the best Π_k amounts to balancing the approximation error $V_\Pi^* - V_{\Pi_k}^*$ and the estimation error $V_{\Pi_k}^* - V(\hat{\pi}_{l,k})$. The estimation error is at the same rate as $\mathbb{E}[\hat{C}_{n,k}]$ under the hold-out penalty (Mbakop and Tabord-Meehan, 2021). Also, intuitively, one could think that the term $\hat{C}_{n,k} = \hat{V}_l^E(\hat{\pi}_{l,k}) - \hat{V}_r^E(\hat{\pi}_{l,k})$ as an estimator for $V_{\Pi_k}^* - V(\hat{\pi}_{l,k})$, or at least a measure of over-fitting. Therefore, the above result shows the oracle property of $\hat{\pi}_n^{PWM}$: it behaves as if we knew the right class ex ante and used it to compute the optimal treatment rule.

---

the l by n here (didn't write $\hat{\pi}_{l,\hat{k}}$) to stress that the rule now depends on the whole sample as $\hat{k}$ depends on the whole sample.

[4] Our result refines Theorem 3.1. and Corollaries 3.2 and 3.3. of Mbakop and Tabord-Meehan (2021) for holdout penalty and a finite number of policy classes.

The difference in $V^*_{\Pi_k} - V(\hat{\pi}_{l,k})$ is in π but the difference in $\hat{V}^E_l(\hat{\pi}_{l,k}) - \hat{V}^E_r(\hat{\pi}_{l,k})$ is in the way how V is estimated, so I don't see why $\hat{V}^E_l(\hat{\pi}_{l,k})$ would be estimating $V^*_{\Pi_k}$ but not $\hat{V}^E_r(\hat{\pi}_{l,k})$. Ideally we should have $\hat{V}^E_l(\pi^*_k)$, then you would say $\hat{\pi}_{l,k}$ is estimating π^*_k, but then what about $\hat{V}^E_r(\hat{\pi}_{l,k})$, why is this not estimating $\hat{V}^E_r(\pi^*_k)$ and then in turn also estimating $V^*_{\Pi_k}$

The goal of this chapter is to construct an estimator with a similar oracle property in a more general setting of Section 3.2 by combining doubly-robust welfare estimator and model selection.

3.4 Main Results

We return to the general setting introduced in Section 3.2. Under Assumption 3.2.1, the welfare can be written as

$$V(\pi) = \mathbb{E}[\pi(X)\Gamma(W)],$$

and the feasible sample analog is given by

$$\hat{V}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\hat{\Gamma}_i.$$

We further require that the estimated orthogonal scores $\hat{\Gamma}_i$ are computed using J-fold cross-fitting, defined as follows. Split the sample into J evenly sized folds of size $\lfloor n/J \rfloor$ distributing the remaining observations uniformly, and let $j : \{1, \dots, n\} \to \{1, \dots, J\}$ be a function that identifies the fold $j(i)$ to which observation i belongs. Then, let $\hat{g}^{(-j(i))}$, $\hat{m}^{(-j(i))}$, and $\tau^{(-j(i))}_{\hat{m}}$ denote the first-stage estimators computed using $(1 - J^{-1})n$ observations excluding the fold $j(i)$, and compute

$$\hat{\Gamma}_i = \tau^{(-j(i))}_{\hat{m}}(X_i, T_i) + \hat{g}^{(-j(i))}(X_i, Z_i)(Y_i - \hat{m}^{(-j(i))}(X_i, T_i)).$$

Following Athey and Wager (2021), we define a Doubly-Robust EWM estimator as

$$\hat{\pi}^{REWM}_n = \operatorname*{argmax}_{\pi \in \Pi} \hat{V}_n(\pi).$$

Our first goal is to bound its expected regret in finite samples. To this end, we define:

$$\tilde{V}_n(\pi) = \frac{1}{n}\sum_{i=1}^{n} \pi(X_i)\Gamma_i,$$

and show that, under Assumption 3.2.1-2 and appropriate moment conditions,

$$\mathbb{E}\left[\sup_{\pi\in\Pi} |\tilde{V}_n(\pi) - V(\pi)|\right] \leqslant \tilde{C}\sqrt{\frac{VC(\Pi)}{n}} \qquad \mathbb{E}\left[\sup_{\pi\in\Pi} |\hat{V}_n(\pi) - \tilde{V}_n(\pi)|\right] = o(n^{-1/2}).$$

That is, not knowing the propensity scores (and other nuisance parameters) only comes at a $o(n^{-1/2})$ price, meaning that $\hat{\pi}_n^{REWM}$ has the optimal rate of convergence.

Here, we impose more explicit restrictions on the distributions of the data, in line with our Assumptions 3.2.2 and 3.2.4. Specifically, we define:[5]

$$\mathcal{P}_{B_\tau,B,\eta} = \left\{ P \in \mathbf{P} : \begin{array}{c} \mathbb{E}_P[\tau_m^2(X,T)] \leqslant B_\tau^2 \\ \mathbb{E}_P[(Y - m(X,T))^2|X,T] \overset{a.s.}{\leqslant} B^2 \\ \sup_{x,z} |g(x,z)| \leqslant \eta^{-1} \end{array} \right\}, \tag{3.2}$$

and prove the following result.

Theorem 3.4 (Doubly-Robust EWM)**.** *Let Assumptions 3.2.1 – 3.2.4 hold and $\hat{\pi}_n^{REWM}$ denote the Doubly-Robust EWM estimator defined above, with the first stage estimators for the nuisance parameters constructed using a J-fold cross-fitting. Then,*

$$\sup_{P\in\mathcal{P}_{B_\tau,B,\eta}} \mathbb{E}_P[R(\hat{\pi}_n^{REWM})] \leqslant C\frac{\sqrt{B_\tau^2\eta^2 + B^2}}{\eta}\sqrt{\frac{VC(\Pi)}{n}} + R_n,$$

where $C \leqslant 58$ is a universal constant, and $R_n = 2(R_{1,n} + R_{2,n} + R_{3,n})$ with

$$R_{1,n} = C\sqrt{(J+2)B^2 \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_g}}},$$

$$R_{2,n} = C\sqrt{(J+2)\frac{2(\eta^2+1)}{\eta^2} \cdot \frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}},$$

$$R_{3,n} = \sqrt{\frac{a((1-J^{-1})n)^2}{n^{\zeta_m+\zeta_g}}}.$$

[5]To relate this with the set $\mathcal{P}_{B,\eta}$ defined prior to Theorem 3.1, recall from Example 3.1 that $\tau_m(X,T) = m(X,1) - m(X,0)$ so that $\mathbb{E}(\tau_m^2) \leqslant 4B^2$ provided that $\mathbb{E}_P[(Y - m(X,T))^2|X,T] \leqslant B^2$. The latter neither implies nor is implied by $\mathbb{E}_P[Y^2] \leqslant B^2$.

It is instructive to compare this result with Theorem 3.1 in the context of binary treatments under unconfoundedness (see Example 3.1). Recall that when the propensity scores are unknown, the analog of Theorem 3.1 holds with an additional $O(\phi_n^{-1})$ term, where ϕ_n is a rate of convergence of the propensity score estimator. The latter is generally slower than root-n, meaning that $\hat{\pi}_n^{EWM}$ is not rate-optimal. On the other hand, under the assumptions of Theorem 3.4, the extra term in the upper bound is $R_n = o(n^{-1/2})$. Therefore, $\hat{\pi}_n^{REWM}$ is rate-optimal, whether the propensity score is known or not, which illustrates the main advantage of using robust welfare estimates.

Next, we present our main result which adds model selection. We propose using a Robust Penalized Welfare Maximization (RPWM) treatment rule, defined as follows.

1. Let $l = \lceil (1-s)n \rceil$ and $r = n - l$ for some $s \in (0,1)$, and call $W_1, \ldots, W_l$ the estimating sample, and $W_{l+1}, \ldots, W_{l+r}$ the test sample. We use subscripts l, r, and n for quantities that depend on the estimating sample only, on the test sample only, and on the entire sample.

2. Compute the RWM rules $\hat{\pi}_{l,k} \equiv \text{argmax}_{\pi \in \Pi_k} \hat{V}_l(\pi)$ for each Π_k using the estimating sample with $\hat{\Gamma}_i$ computed using a J-fold cross-fitting. Evaluate each $\hat{\pi}_{l,k}$ by computing the penalized welfare $Q_{n,k}(\hat{\pi}_{l,k}) = \hat{V}_l(\hat{\pi}_{l,k}) - \hat{C}_{n,k}$ where the penalty is $\hat{C}_{n,k} = \hat{V}_l(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k})$.

3. Select $\hat{k} = \text{argmax}_k Q_{n,k}(\hat{\pi}_{l,k})$, and define $\hat{\pi}_n^{RPWM} \equiv \hat{\pi}_{n,\hat{k}}$.

The following result shows that such estimator automatically selects the best class and attains the optimal rate of convergence.

Theorem 3.5. *Let Assumptions 3.2.1 – 3.2.4 hold and $\hat{\pi}_n^{RPWM}$ denote the Doubly-Robust PWM estimator defined above, with the first stage estimators for the nuisance parameters constructed using a J-fold cross-fitting. Then:*

$$\mathbb{E}_P\left[R(\hat{\pi}_n^{RPWM})\right] \leqslant \inf_{k \leqslant K}\{V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}[\hat{C}_{n,k}]\} + S_n,$$

where V^*_Π *and* $V^*_{\Pi_k}$ *denote the maximum welfare attainable within the corresponding policy classes (both depend on* P*), and* $S_n = O(\sqrt{B_\tau^2\eta^2 + B^2}/\eta \cdot 1/\sqrt{sn})$.

Moreover, letting $\mathcal{P}^k_{B_\tau,B,\eta} \subset \mathcal{P}_{B_\tau,B,\eta}$ *be a set of distributions such that* $V^*_\Pi = V^*_{\Pi_k}$,

$$\sup_{P\in\mathcal{P}^k_{B_\tau,B,\eta}} \mathbb{E}_P[R(\hat{\pi}_n^{RPWM})] \leqslant \frac{\sqrt{B_\tau^2\eta^2 + B^2}}{\eta}\left(C\sqrt{\frac{VC(\Pi_k)}{(1-s)n}} + K\sqrt{\frac{1}{sn}}\right) + S^k_{1,n} + S_{2,n}$$

where $C \leqslant 58$ *is a universal constant,* $S^k_{1,n} = R^k_{1,(1-s)n} + R^k_{2,(1-s)n} + R_{3,(1-s)n}$, *where* $R^k_{1,n}$, $R^k_{2,n}$, *and* $R_{3,n}$ *are given in Theorem 3.4 with* Π_k *instead of* Π, *and* $S_{2,n} = o(n^{-1/2})$.

Note that this theorem is comparable to Theorem 3.3. It shows the same oracle property as PWM discussed in Section 3.3. Moreover, by incorporating the doubly robust score, RPWM can retain the $n^{-1/2}$ rate in more general settings as the REWM rule. Hence, we are able to get the benefit of both worlds.

3.5 Simulation

In this section, we conduct a simple simulation to demonstrate how RPWM rule balances between approximation error and estimation error.

We generate a random sample of size n with the following DGP.

$$\begin{aligned} Y(0) &= 0.7(X_3 + X_4 + \epsilon_0), \\ Y(1) &= X_2 - X_1 + 0.7(X_3 + X_4 + \epsilon_1), \\ P(T = 1|X) &= \Lambda(\log(0.5) + (X_1 + X_2 + X_3 + X_4)(\log(2) - \log(0.5))/4). \end{aligned}$$

where all covariates follow $U[0,1]$ and errors follow $N(0,1)$ independently. The $\Lambda(\cdot)$ denotes the logistic function so the propensity score is in between $\frac{1}{3}$ to $\frac{2}{3}$. Under this DGP, the average treatment effect is zero. However, There is heterogeneous treatment effect and

$$\mathbb{E}[Y(1) - Y(0)|X] = X_2 - X_1,$$

which suggest that the first best treatment policy is $\mathbb{1}\{X_2 \geq X_1\}$. Consider the $[0,1]\times[0,1] \subset R^2$ square where (X_1, X_2) belongs, the 45 degree diagonal line across this square would be the boundary of the first best treatment policy.

Now, for policy rule learning, suppose we arbitrarily decided to focus on decision trees that only splits on X_1 and X_2 and up to depth 4. We compare three different algorithms, the first one only considers depth 2 trees, the second one only considers depth 4 trees, and then an adaptive one which chooses across depths 2, 3 and 4 using the hold-out penalty. The last algorithm corresponds to RPWM and the first two REWM. We run Monte Carlo simulations with $n \in \{200, 400, 800, 1200, 1600, 2000\}$ and plot the regrets in Figure 3.1. 200 simulations were run for each sample size.

We see that when sample size is small, the estimation error would dominate, hence focusing on depth 2 trees leads to less regret. When the sample size is large, the approximation error would dominate so depth 4 trees become more favorable. The adaptive RPWM rule should ideally trace the lower envelope of the other two curves. That is similar to what it behaves in this simulation. We do notice a relatively poorer performance when sample size is small. This might be due to the fact that hold-out penalty effectively reduce sample size.

At last, we show some policy rules learned from the depth 2 and 4 trees at $n = 200$ and 2000 in Figure 3.2. We can see that the depth 4 tree behaves poorly at $n = 200$ due to over-fitting while does a good job approximating the first best policy rule when $n = 2000$.

3.6 Conclusion

In this chapter, we studied model selection in doubly robust policy learning. Following Mbakop and Tabord-Meehan (2021) and Athey and Wager (2021), we added hold-out penalty to the doubly robust policy learning algorithm. The resulting method could achieve data-driven model selection while retaining optimal $n^{-1/2}$ rate under general setups including quasi-experiments where propensity scores are unknown. By deriving finite sample upper

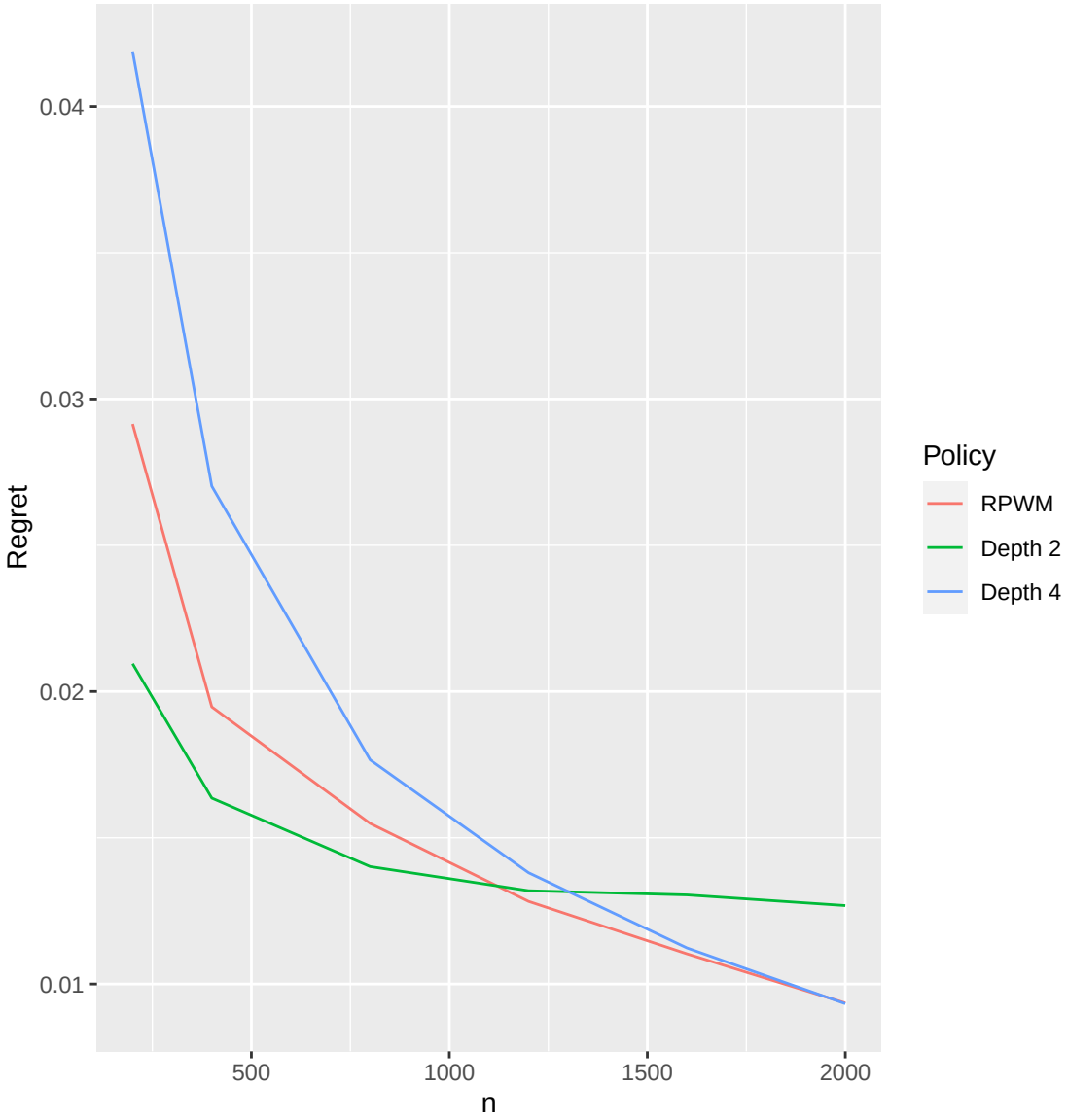
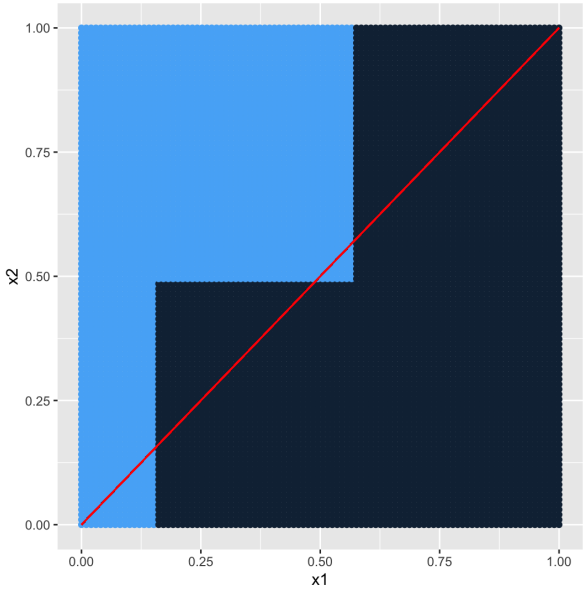
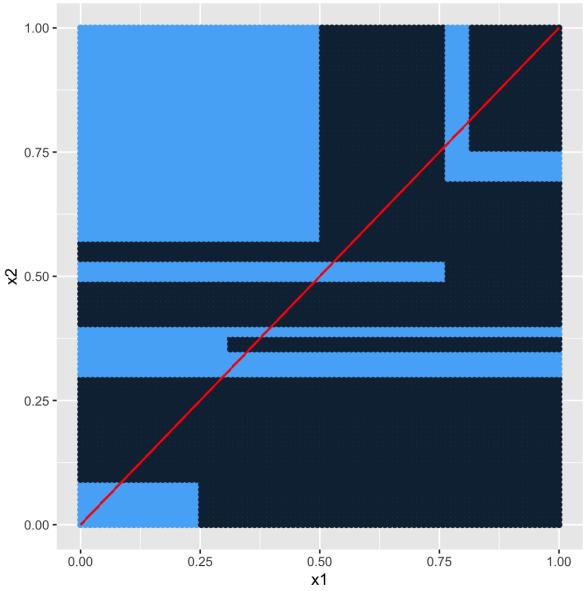


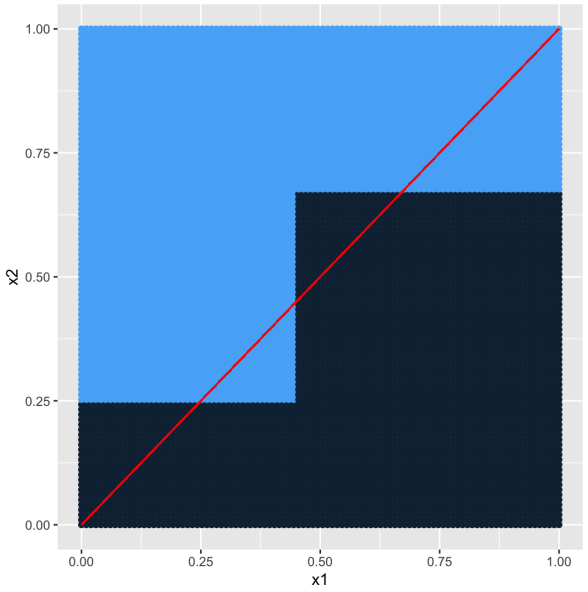
Figure 3.1: Regrets of 3 Algorithms with Different Sample Sizes



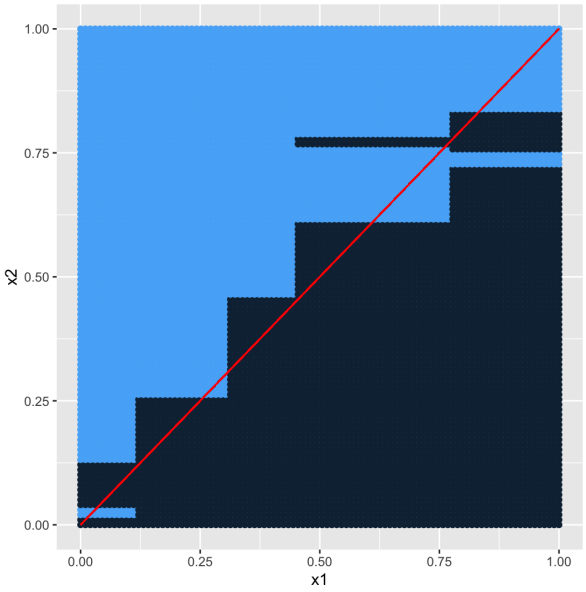
(a) A depth 2 tree with $n = 200$.



(b) A depth 4 tree with $n = 200$.



(c) A depth 2 tree with $n = 2000$.



(d) A depth 4 tree with $n = 2000$.

Figure 3.2: Examples of Policy Trees Learned with $n = 200$ and 2000.

bounds on expected regret, we show that the algorithm can automatically balance approximation error with estimation error. We also refined some related results in the literature and derived a new finite sample lower bound to show that the $n^{-1/2}$ rate is indeed optimal.

3.7 Appendix

3.7.1 Known Results for Reference and Some Refinements

First, we recite a well-known symmetrization inequality. See, e.g., Lemma 2.3.1. in van der Vaart and Wellner (1996).

Lemma 3.1 (Symmetrization). *Let $W_1, \ldots, W_n$ be an i.i.d. sample. Then for any class of measurable functions $\mathcal{F}$,*

$$\mathbb{E}\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n} f(W_i) - \mathbb{E}(f(W_i))\right|\right] \leqslant 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i f(W_i)\right|\right]$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables independent from $W_1, \ldots, W_n$.

Let ψ be a strictly increasing, convex function with $\psi(0) = 0$ and X be a random variable. Then the Orlisz norm $||X||_\psi$ is defined as

$$||X||_\psi = \inf\left\{C > 0 : \mathbb{E}\left(\psi\left(\frac{|X|}{C}\right)\right) \leqslant 1\right\}.$$

Then, the following maximal inequality holds.

Lemma 3.2 (Maximal Inequality with Orlisz Norms). *For any random variables $X_1, \ldots, X_n$ and any strictly increasing, convex function ψ,*

$$\mathbb{E}\left[\max_{j\leqslant m}|X_j|\right] \leqslant \psi^{-1}(m)\max_{j\leqslant m}||X_j||_\psi$$

Proof. For any $C > 0$,

$$\begin{aligned}\psi\left(\mathbb{E}\left[\max_{j\leqslant m}\tfrac{|X_j|}{C}\right]\right) &\leqslant \mathbb{E}\left[\max_{j\leqslant m}\psi\left(\tfrac{|X_j|}{C}\right)\right] \\ &\leqslant m\max_{j\leqslant m}\mathbb{E}\left[\psi\left(\tfrac{|X_j|}{C}\right)\right],\end{aligned}$$

where the first inequality holds because ψ is convex and non-decreasing. Therefore, for any C such that $\max_{j\leqslant m}\mathbb{E}\left[\psi\left(|X_j|/C\right)\right]\leqslant 1$, we have

$$\mathbb{E}\left[\max_{j\leqslant m}|X_j|\right]\leqslant C\psi^{-1}(m).$$

Choosing $C=\max_{j\leqslant m}||X_j||_\psi$ concludes the proof. ■

The following result is Theorem 2.6.4. from Van der Vaart and Wellner (1996) with a precisely pineed down universal constant.

Lemma 3.3 (Covering Numbers for VC classes)**.** *For any VC-class $\mathcal{C}$ of sets, any probability measure Q, any $r\geqslant 1$, and $0<\varepsilon<1$,*

$$N(\varepsilon,\mathcal{C},L_r(Q))\leqslant\frac{1}{2\sqrt{e}}V(\mathcal{C})(4e)^{V(\mathcal{C})}\left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{C})-1)}.$$

Proof. We closely follow the proof of Theorem 2.6.4. in van der Vaart and Wellner (1996). We start by referencing the main steps and introducing the necessary notation. First, note that $||\mathbf{1}_C-\mathbf{1}_D||_{Q,r}=Q^{1/r}(C\triangle D)$, so an ε^r-cover under $L_1(Q)$ produces an ε-cover under $L_r(Q)$. Therefore, the result for $r>1$ follows immediately from the result for $r=1$. Second, one can argue that it suffices to consider empirical type measures Q supported on a large enough finite set of distinct points $\{x_1,\ldots,x_n\}$. Third, it is more convenient to bound the packing number $D(\varepsilon,\mathcal{C},L_1(Q))$ first and use the fact that $N(\varepsilon,\mathcal{C},L_1)(Q))\leqslant D(\varepsilon/2,\mathcal{C},L_1(Q))$.

Each set $C\in\mathcal{C}$ can be identified with a binary vector $\mathbf{1}_C=(\mathbf{1}(x_i\in C))_{i=1}^n$, and the collection $\mathcal{C}$ can be identified with a binary matrix $\mathcal{Z}$ of size $n\times\#\mathcal{Z}$. Define $d(\mathbf{1}_{C_1},\mathbf{1}_{C_2})=n^{-1}\sum_{i=1}^n|\mathbf{1}_{C_1}-\mathbf{1}_{C_2}|$. Then, recalling that Q places probability $1/n$ on each x_i, $Q(C_1\triangle C_2)=d(\mathbf{1}_{C_1},\mathbf{1}_{C_2})$, so that $D(\varepsilon,\mathcal{C},L_1(Q))=D(\varepsilon,\mathcal{Z},d)$. For simplicity of notation, assume that $\mathcal{Z}$ is ε-separated with respect to d, so the goal is to bound its size $\#\mathcal{Z}$ in terms of the VC dimension $V(\mathcal{C})$.

Denote $S=V(\mathcal{C})-1$ and fix an integer m such that $S\leqslant m<n$. For a subset $J\subset\{1,\ldots,n\}$ of size $\#J=m$, let $\mathcal{Z}_J$ denote the projection of $\mathcal{Z}$ onto $\{0,1\}^J$, and $\overline{\#\mathcal{Z}_J}$

denote the average size of $\mathcal{Z}_J$ over all subsets J or size m. Then, following the proof on Page 138 of van der Vaart and Wellner (1996) we arrive to the bound

$$\#\mathcal{Z} \leqslant \frac{\overline{\#\mathcal{Z}_J} n\varepsilon(m+1)}{\varepsilon n(m+1) - 2(n-m)S} \leqslant \frac{\varepsilon(m+1)\overline{\#\mathcal{Z}_J}}{\varepsilon(m+1) - 2S} \leqslant \frac{\varepsilon m \overline{\#\mathcal{Z}_J}}{\varepsilon m - 2S},$$

which holds without any extra constants. The number of points in any $\mathcal{Z}_J$ is equal to the number of subsets picked out by $\mathcal{C}$ from the points $\{x_i : i \in J\}$. By the Sauer-Shelah Lemma, this is bounded by $\sum_{j=0}^{S} \binom{m}{j}$, which is smaller than $(em/S)^S$ for $m \geqslant S$.[6] Therefore,

$$\#\mathcal{Z} \leqslant \left(\frac{e}{S}\right)^S \frac{m^{S+1}\varepsilon}{m\varepsilon - 2S}$$

holds for all integers m such that $S \leqslant m < n$. Denote the right-hand side of the preceding display by $f(m)$. This function is strictly decreasing until $m^* = 2(S+1)/\varepsilon$ and strictly increasing afterwards. Therefore, the optimal unconstrained choice is $m = m^*$, for which $f(m^*) = (2e/\varepsilon)^S (S+1)(1+S^{-1})^S$. However, the argument leading to the upper bound on $\#\mathcal{Z}$ only applies to integer m such that $S \leqslant m < n$. To ensure that a similar bound holds for an integer value of m, we can simply use $f(m^*-1)$ since somewhere between m^*-1 and m^* there must be an integer, and $f(m)$ is decreasing on this interval. We have

$$\begin{aligned}
f(m^*-1) &= \left(\tfrac{e}{S}\right)^S \frac{(2(S+1)/\varepsilon - 1)^{S+1}\varepsilon}{(2(S+1)/\varepsilon - 1)\varepsilon - 2S} \\
&= \left(\tfrac{2e}{\varepsilon}\right)^S \tfrac{1}{1-\varepsilon/2}(S+1-\varepsilon/2)\left(1 + \tfrac{1-\varepsilon/2}{S}\right)^S \\
&\leqslant \left(\tfrac{2e}{\varepsilon}\right)^S (S+1)\tfrac{1}{1-\varepsilon/2}\exp(1-\varepsilon/2) \\
&\leqslant \left(\tfrac{2e}{\varepsilon}\right)^S (S+1)\cdot 2\sqrt{e},
\end{aligned}$$

for all $\varepsilon \in (0,1)$ since the function $g(\varepsilon) = (1-\varepsilon/2)^{-1}\exp(1-\varepsilon/2)$ is monotonically increasing. Therefore, we obtain the bound

$$\#\mathcal{Z} \leqslant \left(\frac{2e}{\varepsilon}\right)^S (S+1)\cdot 2\sqrt{e},$$

[6] Indeed, for $t \in (0,1)$, $\sum_{j=0}^{S}\binom{m}{j} \leqslant \sum_{j=0}^{S}\binom{m}{j}\frac{t^j}{t^S} \leqslant \frac{(1+t)^m}{t^s}$. Set $t = \frac{S}{m}$ and use $(1+S/m)^m \leqslant e^S$.

and it remains to check that this bound still holds when $m^* - 1 < S$ or $m^* \geqslant n$. Note that $m^* - 1 \geqslant S$ for all $\varepsilon \in (0,1)$. If $m^* \geqslant n$, by the Sauer-Shelah Lemma

$$\#\mathcal{Z} \leqslant \sum_{j=0}^{S} \binom{n}{j} \leqslant \left(\frac{en}{S}\right)^S \leqslant \left(\frac{em^*}{S}\right)^S \leqslant e\left(\frac{2e}{\varepsilon}\right)^S,$$

which certainly implies the bound in the previous display. Therefore, recalling that $\#\mathcal{Z} = D(\varepsilon, \mathcal{C}, L_1(Q))$,

$$\begin{aligned} N(\varepsilon, \mathcal{C}, L_1(Q)) &\leqslant D(\varepsilon/2, \mathcal{C}, L_1(Q)) \\ &\leqslant \left(\tfrac{4e}{\varepsilon}\right)^S (S+1) \cdot 2\sqrt{e} \\ &= \left(\tfrac{4e}{\varepsilon}\right)^{V(\mathcal{C})-1} V(\mathcal{C}) \cdot 2\sqrt{e} \\ &= \tfrac{1}{2\sqrt{e}} V(\mathcal{C})(4e)^{V(\mathcal{C})} \left(\tfrac{1}{\varepsilon}\right)^{(V(\mathcal{C})-1)}, \end{aligned}$$

and the desired result follows.

■

Next, we state and prove two simple lemmas about a specific VC-subgraph class of functions. A subgraph of a function $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$C_f = \{(t, x) \in \mathbb{R} \times \mathcal{X} : t < f(x)\}.$$

A class of functions $\mathcal{F}$ is VC-subgraph if the class of all subgraphs

$$\mathcal{C}_{\mathcal{F}} = \{C_f : f \in \mathcal{F}\}$$

has a finite VC dimension. In this case we denote $V(\mathcal{F}) = V(\mathcal{C}_{\mathcal{F}})$.

The next result is Theorem 2.6.7. from van der Vaart and Wellner (1996). It is a direct consequence of the result for sets and holds with the same universal constant.

Lemma 3.4 (Covering Number for VC-subgraph Classes)**.** *For a VC-class of functions with measurable envelope function F and $r \geqslant 1$, one has for any probability measure Q with $||F||_{Q,r} > 0$,*

$$N(\varepsilon ||F||_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)},$$

for $0 < \varepsilon < 1$.

Next, we refine the above result for a particular VC-subgraph class of functions.

Lemma 3.5 (A Simple VC-Subgraph Class)**.** *Let $\mathcal{G}$ denote a class of subsets of $\mathcal{X}$ with a finite VC dimension $V(\mathcal{G})$, and $F : \mathcal{X} \to \mathbb{R}$ be an arbitrary function. Define a class of functions:*

$$\mathcal{F} = \{\mathbf{1}(x \in G)F(x) : G \in \mathcal{G}\}.$$

Then, $\mathcal{F}$ is VC-subgraph with $V(\mathcal{F}) \leqslant V(\mathcal{G})$.

Proof. Let $VC(\mathcal{G}) = d$ and $D = \{(t_1, x_1), \ldots, (t_d, x_{d+1})\} \subset \mathbb{R} \times \mathcal{X}$ be an arbitrary set of points. By definition, D is shattered by $\mathcal{F}$ if for every subset $\{(t_j, x_j) : j \in J\}$ there is a function f with subgraph C_f such that $C_f \cap D = \{(t_j, x_j) : j \in J\}$. Equivalently, D is shattered by $\mathcal{F}$ if for every subset $J \subset \{1, \ldots, d+1\}$ there is a set $G \in \mathcal{G}$ satisfying

$$\begin{aligned} t_j &< \mathbf{1}(x_j \in G)F(x_j) \text{ for } j \in J \\ t_k &\geqslant \mathbf{1}(x_k \in G)F(x_k) \text{ for } k \notin J \end{aligned} \tag{3.3}$$

We will argue that D cannot be shattered by $\mathcal{F}$.

First, if there is (t_j, x_j) such that $t_j < 0$ and $t_j < F(x_j)$, then $t_j < \mathbf{1}(x_j \in G)F(x_j)$ holds for all $G \in \mathcal{G}$. In this case, any subset of D that does not include t_j, x_j cannot be picked out, so D cannot be shattered by $\mathcal{F}$. Similarly, if there is (t_k, x_k) such that $t_k \geqslant 0$ and $t_k \geqslant F(x_k)$, then $t_k \geqslant \mathbf{1}(x_k \in G)F(x_k)$ holds for all $G \in \mathcal{G}$. So, any subset of D that includes this point cannot be picked out and D cannot be shattered by $\mathcal{F}$. Therefore, we will assume that each (t_j, x_j) satisfies either $t_j < 0, F(x_j) \geqslant 0$ or $t_j \geqslant 0, F(x_j) < 0$ for $j = 1, \ldots, d+1$.

Recall that $\mathcal{G}$ does not shatter $\{x_1, \ldots, x_{d+1}\}$, meaning that there exist a subset $\{x_j\}_{j \in J}$ that $\mathcal{G}$ cannot pick out. Then, for every $G \in \mathcal{G}$ we have either $x_j \notin G$ for some $j \in J$ or $x_k \in G$ for some $k \notin J$. If the inequalities in (3.3) do not hold for this J for any G, then $\{(t_j, x_j)\}_{j \in J}$ cannot be picked out and D cannot be shattered by $\mathcal{F}$. Suppose the inequalities in (3.3) hold for some $G \in \mathcal{G}$. If $x_j \notin G$ for some $j \in J$, it must be that $t_j < 0$ and, according to the previous discussion, $F(x_j) \geqslant 0$. Then the set $J' = J \backslash (t_j, x_j)$ cannot be picked out. If

$x_k \in G$ for some $k \notin J$, it must be that $t_k \geqslant 0$ and $F(x_k) < 0$, so the set $J'' = J \cup k$ cannot be picked out. Therefore, D cannot be shattered by $\mathcal{F}$ and $VC(\mathcal{F}) \leqslant VC(\mathcal{G})$.

■

Lemma 3.6 (Covering Numbers for Special VC-Subgraph Classes)**.** *Let $\mathcal{G}$ denote a class of subsets of $\mathcal{X}$ with a finite VC dimension $V(\mathcal{G})$, and $F : \mathcal{X} \to \mathbb{R}$ be an arbitrary function. Define a class of functions:*

$$\mathcal{F} = \{\mathbf{1}(x \in G)F(x) : G \in \mathcal{G}\}.$$

Then, for any $r \geqslant 1$, probability measure Q with $||F||_{Q,r} > 0$, and $0 < \varepsilon < 1$,

$$N(\varepsilon ||F||_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}.$$

Proof. By Lemma 3.5, $\mathcal{F}$ is VC-subgraph. For $r = 1$, note that:

$$||f_1 - f_2||_{Q,1} = \mathbb{E}_Q[|\mathbf{1}_{G_1} - \mathbf{1}_{G_2}||F|] = P(C_{f_1} \triangle C_{f_2}) ||F||_{Q,1},$$

where $P = \lambda \times Q / ||F||_{Q,1}$ is a probability measure on $\mathbb{R} \times \mathcal{X}$ and λ is a Lebesgue measure on $\mathbb{R}$. Then, by Lemma 3.3,

$$N(\varepsilon ||F||_{Q,1}, \mathcal{F}, L_1(Q)) = N(\varepsilon, \mathcal{C}_{\mathcal{F}}, L_1(P)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{(V(\mathcal{F})-1)}.$$

For $r > 1$, note that:

$$||f_1 - f_2||_{Q,r}^r = \mathbb{E}_Q(|\mathbf{1}_{G_1}F - \mathbf{1}_{G_2}F||F|^{r-1}) = \frac{||f_1 - f_2||_{R,1}}{||F||_{R,1}} \mathbb{E}_Q(|F|^r),$$

for the probability measure R with density $|F|^{r-1}/\mathbb{E}_Q(|F|^{r-1})$ with respect to Q. Therefore,

$$||f_1 - f_2||_{Q,r} = \left(\frac{||f_1 - f_2||_{R,1}}{||F||_{R,1}}\right)^{1/r} ||F||_{Q,r},$$

so that by the previous argument applied to R instead of Q

$$N(\varepsilon ||F||_{Q,r}, \mathcal{F}, L_r(Q)) \leqslant N(\varepsilon^r ||F||_{R,1}, \mathcal{F}, L_1(R)) \leqslant \frac{1}{2\sqrt{e}} V(\mathcal{F})(4e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)}$$

■

3.7.2 Auxiliary Lemmas

Now we are ready to state and prove three auxiliary lemmas that give our main results.

Lemma 3.7 (Finite-Sample Bound on Rademacher Complexity)**.** *Let $W_1, \ldots, W_n$ be an i.i.d. sample and $\xi_1, \ldots, \xi_n$ be i.i.d. Rademacher random variables independent of $W_1, \ldots, W_n$.*

1. *Let $\mathcal{F}$ be a VC-subgraph of functions with $f_0(w) = 0 \in \mathcal{F}$, a finite VC dimension $VC(\mathcal{F})$, and a measurable envelope F such that $S = \mathbb{E}(F^2) < \infty$. Then:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i f(W_i)\right|\right] \leqslant C\sqrt{\frac{VC(\mathcal{F})S}{n}},$$

   where $C = 4\sqrt{12}\int_0^1 \sqrt{1/(2e^{3/2}) + \log(16e) + 2\log(1/u)}du \leqslant 34$.

2. *In the special case when $\mathcal{F} = \{f(x) = 1(x \in G)F(x) : G \in \mathcal{G}\}$, for a VC-class of sets $\mathcal{G}$ and an arbitrary measurable function F with $S = \mathbb{E}(F^2) < \infty$, the above holds with $C = 4\sqrt{12}\int_0^1 \sqrt{1/(2e^{3/2}) + \log(4e) + 2\log(1/u)}du \leqslant 29$.*

Proof. Denote $\mathbb{G}_n^0(f) = n^{-1/2}\sum_{i=1}^n \xi_i f(W_i)$. By the Law of Iterated Expectations,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\mathbb{G}^0(f)\right|\right] = \frac{1}{\sqrt{n}}\mathbb{E}_{W_1^n}\left[\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{G}^0(f)\right|\right]\right] \tag{3.4}$$

We will use a simple chaining argument to bound the right hand side of (3.4). Let $\eta = 2\,||F||_{2,n}$, and define $\mathcal{F}_0 = \{f_0\}$ and $\mathcal{F}_j$ contain centers of the balls in the minimal $\eta 2^{-j}$-cover of $\mathcal{F}$ under $||\cdot||_{2,n}$, so that $|\mathcal{F}_j| = N(\eta 2^{-j}, \mathcal{F}, ||\cdot||_{2,n})$. Let $\phi_j : \mathcal{F} \to \mathcal{F}_j$ be a map that for a given f finds the closest element of $\mathcal{F}_j$. For any $f_k \in \mathcal{F}_k$ define a chain $f_{k-l} = \phi_{k-l}(f_{k-l+1})$ for $l = 1, \ldots, k$. Then,

$$\mathbb{G}_n^0(f_k) = \sum_{j=1}^{k}(\mathbb{G}_n^0(f_j) - \mathbb{G}_n^0(f_{j-1})) \leqslant \sum_{j=1}^{k}\max_{g\in\mathcal{F}_j}|\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))|,$$

Let $\psi_2(x) = e^{x^2} - 1$ and $||\cdot||_{\psi_2}$ denote the corresponding Orlisz norm. By Lemma 2.2.7. in van der Vaart and Wellner (1996), conditional on W_1^n, the process $\mathbb{G}_n^0(f)$ is sub-Gaussian

for the metric $d_n(f_1, f_2) = ||f_1 - f_2||_{2,n}$, and satisfies $||\mathbb{G}_n^0(f) - \mathbb{G}_n^0(g)||_{\psi_2} \leqslant \sqrt{6}\,||f - g||_{2,n}$. By Lemma 3.2 and the above discussion,

$$\begin{aligned}\mathbb{E}_{\xi_1^n}\left[\max_{g\in\mathcal{F}_j}|\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))|\right] &\leqslant \psi_2^{-1}(|\mathcal{F}_j|)\max_{g\in\mathcal{F}_j}||\mathbb{G}_n^0(g) - \mathbb{G}_n^0(\phi_{j-1}(g))||_{\psi_2} \\ &\leqslant \sqrt{6}\cdot\psi_2^{-1}(N(\eta 2^{-j}, \mathcal{F}, ||\cdot||_{2,n}))\cdot\eta 2^{-(j-1)}\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}_k}|\mathbb{G}_n^0(f)|\right] &\leqslant \sqrt{6}\sum_{j=1}^{k}\psi_2^{-1}(N(\eta 2^{-j}, \mathcal{F}, ||\cdot||_{2,n}))\eta 2^{-(j-1)} \\ &\overset{(a)}{\leqslant} 4\sqrt{6}\int_0^{\eta/2}\psi^{-1}(N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n}))d\varepsilon \\ &= 4\sqrt{6}\int_0^{||F||_{2,n}}\sqrt{\log(N(\varepsilon, \mathcal{F}, ||F||_{2,n}) + 1)}d\varepsilon \\ &\overset{(b)}{\leqslant} 4\sqrt{12}\int_0^{||\cdot||_{2,n}}\sqrt{\log N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n})}d\varepsilon,\end{aligned}$$

where (a) follows from rearranging rectangles under the curve $\varepsilon \mapsto \psi_2^{-1}(N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n}))$, and (b) follows from $\log(x+1) \leqslant 2\log(x)$ for $x \geqslant 2$. Since, conditional on W_1^n, the process $\mathbb{G}_n^0$ is separable, by letting $k \to \infty$ in the previous display we conclude that

$$\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n^0(f)|\right] \leqslant 4\sqrt{12}\int_0^{||F||_{2,n}}\sqrt{\log N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n})}d\varepsilon. \tag{3.5}$$

Denote $V \equiv VC(\mathcal{F})$ and $K = (2\sqrt{e})^{-1}$. Applying Lemma 3.4 (or Lemma 3.6 for the special case) with $r = 2$ and $Q = P_n$,

$$\begin{aligned}\log N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n}) &\leqslant \log(KV) + V\log(16e) + 2(V-1)\log\left(\tfrac{||F||_{2,n}}{\varepsilon}\right) \\ &= V\left(K^{\frac{\log(KV)}{KV}} + \log(16e) + 2\tfrac{V-1}{V}\log\left(\tfrac{||F||_{2,n}}{\varepsilon}\right)\right) \\ &\leqslant V\left(K/e + \log(16e) + 2\log\left(\tfrac{||F||_{2,n}}{\varepsilon}\right)\right),\end{aligned}$$

where the last line uses the fact that $\log(t)/t \leqslant 1/e$ for all $t > 0$. Therefore,

$$\begin{aligned}\int_0^{||F||_{2,n}} \sqrt{\log N(\varepsilon, \mathcal{F}, ||\cdot||_{2,n})}d\varepsilon &\leqslant \int_0^{||F||_{2,n}} \sqrt{K/e + \log(16e) + 2\log\left(||F||_{2,n}/\varepsilon\right)}d\varepsilon \cdot \sqrt{V} \\ &\leqslant \int_0^1 \sqrt{K/e + \log(16e) + 2\log(1/u)}du\sqrt{V\,||F||_{2,n}^2},\end{aligned} \tag{3.6}$$

where the second line follows from a change of variables $u = \varepsilon/\,||F||_{2,n}$. Combining (3.5) and (3.6), we obtain

$$\mathbb{E}_{\xi_1^n}\left[\sup_{f\in\mathcal{F}}|\mathbb{G}_n^0(f)|\right] \leqslant C\sqrt{V\,||F||_{2,n}^2}$$

where $C = 4\sqrt{12}\int_0^1 \sqrt{K/e + \log(16e) + 2\log(1/u)}du$ (or the same expression with $4e$ instead of $16e$ in the special case). By (3.4) and Jensen's inequality,

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i f(W_i)\right|\right] \leqslant C\sqrt{\frac{VC(\mathcal{F})S}{n}},$$

which concludes the proof.

■

3.7.3 Proofs of Theorems 3.1, 3.2, and 3.3

3.7.3.1 Proof of Theorem 3.1

To keep notation simple, we write $\hat{\pi}_n$ instead of $\hat{\pi}_n^{EWM}$. Let π^* denote a rule such that $V(\pi^*) = V_\Pi^* = \sup_{\pi\in\Pi} V(\pi)$. Note that

$$\begin{aligned}R(\hat{\pi}_n) &= V(\pi^*) - V(\hat{\pi}_n) \\ &= V(\pi^*) - \hat{V}_n(\hat{\pi}_n) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n) \\ &\leqslant V(\pi^*) - \hat{V}_n(\pi^*) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n),\end{aligned}$$

and, therefore,

$$\mathbb{E}[R(\hat{\pi}_n)] = \mathbb{E}[\hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n)] \leqslant \mathbb{E}[\sup_{\pi\in\Pi}|\hat{V}_n(\pi) - V(\pi)|].$$

Define a class of functions

$$\mathcal{F} = \left\{ f(w) = \pi(x) \left(\frac{yt}{e(x)} - \frac{y(1-t)}{1-e(x)} \right) : \pi \in \Pi \right\},$$

so that

$$\sup_{\pi \in \Pi} |\hat{V}_n(\pi) - V(\pi)| = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(W_i) - \mathbb{E}[f(W_i)] \right|.$$

Applying Lemma 3.1 and part 2 of Lemma 3.7,

$$\mathbb{E}\left[\left| \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(W_i) - \mathbb{E}[f(W_i)] \right| \right] \leqslant 2C \sqrt{\frac{VC(\mathcal{F})S}{n}},$$

where $C \leqslant 29$ is a universal constant and $S = \mathbb{E}[f(W)^2]$. By Lemma 3.5, $VC(\mathcal{F}) \leqslant VC(\Pi)$, and for any $P \in \mathcal{P}_{B,\eta}$,

$$\mathbb{E}_P[f(W)^2] \leqslant \mathbb{E}_P\left[\frac{Y^2 T}{e(X)^2} + \frac{Y^2(1-T)}{(1-e(X))^2} \right] \leqslant \frac{B^2}{\eta^2},$$

so the desired result follows.

3.7.3.2 Proof of Theorem 3.3

To keep notation simple, we write $\hat{\pi}_n = \hat{\pi}_{n,\hat{k}}$ instead of $\hat{\pi}_n^{PWM}$, and $\hat{V}_n$ instead of $\hat{V}_n^E$. The subscripts n, l, and r, indicate the the corresponding objects depend on the entire sample, only the estimating sample, and only the test sample correspondingly. For example, while $\hat{\pi}_{l,k}$ depends only on the estimating sample, $\hat{\pi}_{n,\hat{k}}$ depends on the entire sample by the choice of $\hat{k}$. Let π_k^* denote a rule such that $V(\pi_k^*) = V_{\Pi_k}^* = \max_{\pi \in \Pi_k} V(\pi)$. Recall that, by definition,

$$Q_{n,k}(\hat{\pi}_{n,\hat{k}}) = \hat{V}_l(\hat{\pi}_{n,\hat{k}}) - \hat{C}_{n,\hat{k}} = \hat{V}_r(\hat{\pi}_{n,\hat{k}}).$$

Write

$$\begin{aligned} R(\hat{\pi}_n) &= V_{\Pi}^* - V_{\Pi_k}^* \\ &\quad + V(\pi_k^*) - Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) \\ &\quad + Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}}) \end{aligned}$$

By the definitions of $\hat{k}$ and $\hat{\pi}_{l,k}$, for any k,

$$V(\pi_k^*) - Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) \leqslant V(\pi_k^*) - Q_{n,k}(\hat{\pi}_{l,k}) \leqslant V(\pi_k^*) - \hat{V}_l(\pi_k^*) + \hat{C}_{n,k},$$

so that

$$\mathbb{E}[V(\hat{\pi}_k^*) - Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}})] \leqslant \mathbb{E}[\hat{C}_{n,k}].$$

Next, write

$$\mathbb{E}[\hat{V}_r(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})] \leqslant r^{-1/2}\mathbb{E}\left[\max_{k\leqslant K}\sqrt{r}|\hat{V}_r(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})|\right], \tag{3.7}$$

and, working conditional on the estimating sample W_1^l,

$$\mathbb{E}\left[\max_{k\leqslant K}\sqrt{r}|\hat{V}_r(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})| \;\middle|\; W_1^l\right] \leqslant K\max_{k\leqslant K}\mathbb{E}\left[\sqrt{r}|\hat{V}_r(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})| \;\middle|\; W_1^l\right]. \tag{3.8}$$

Denoting $f_k(w) = \hat{\pi}_{m,k}(x)(yt/e(x) - y(1-t)/(1-e(x)))$, we have:

$$\begin{aligned}
\mathbb{E}\left[\sqrt{r}|\hat{V}_r(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})| \;\middle|\; W_1^l\right] &= \mathbb{E}\left[\left|r^{-1/2}\textstyle\sum_j f_k(W_j) - \mathbb{E}[f_k(W_j)]\right| \;\middle|\; W_1^l\right] \\
&\leqslant \mathbb{E}\left[\left(r^{-1/2}\textstyle\sum_j f_k(W_j) - \mathbb{E}[f_k(W_j)]\right)^2 \;\middle|\; W_1^l\right]^{1/2} \\
&\leqslant \mathbb{E}[f_k(W_j)^2 \mid W_1^l]^{1/2} \\
&\leqslant \tfrac{B}{\eta},
\end{aligned}$$

where the last inequality follows in the same fashion as in Theorem 3.1. Since this bound does not depend on k, taking expectations on both sides of (3.8) and recalling that $r = ln$, we obtain:

$$\mathbb{E}[\hat{V}_r(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})] \leqslant \frac{B}{\eta}\frac{K}{\sqrt{ln}}.$$

Combining the above results, we conclude that

$$\mathbb{E}[R(\hat{\pi}_n)] \leqslant V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}[\hat{C}_{n,k}] + \frac{B}{\eta}\frac{K}{\sqrt{ln}}, \tag{3.9}$$

holds for all $k \leqslant K$, so that

$$\mathbb{E}[R(\hat{\pi}_n)] \leqslant \inf_{k \leqslant K} \{V^*_\Pi - V^*_{\Pi_k} + \mathbb{E}[\hat{C}_{n,k}]\} + \frac{B}{\eta}\frac{K}{\sqrt{ln}},$$

and the first part of the statement follows.

For the second part of the statement, note that by the Law of Iterated Expectations

$$\begin{aligned}\mathbb{E}[\hat{C}_{n,k}] &= \mathbb{E}[\hat{V}_l(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k}) + V(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k})] \\ &= \mathbb{E}[\hat{V}_l(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})].\end{aligned}$$

Then, repeating the proof of Theorem 3.1 with Π_k instead of Π and m instead of n, we obtain

$$\mathbb{E}[\hat{C}_{n,k}] \leqslant C\frac{B}{\eta}\sqrt{\frac{VC(\Pi_k)}{(1-s)n}}.$$

Plugging this in Equation (3.9) and recalling that $V^*_\Pi = V^*_{\Pi_k}$ for all $P \in \mathcal{P}^k_{B,\eta}$, we conclude that

$$\sup_{P \in \mathcal{P}^k_{B,\eta}} \mathbb{E}_P[R(\hat{\pi}^{PWM}_n)] \leqslant \frac{B}{\eta}\left(C\sqrt{\frac{VC(\Pi_k)}{(1-s)n}} + K\sqrt{\frac{1}{ln}}\right),$$

and the proof is complete.

3.7.3.3 Proof of Theorem 3.2

We consider a particular subclass of $\mathcal{P}_{B,\eta}$ for which the worst-case regret can be bounded from below by a term proportional to $B/\eta\sqrt{d/n}$. The construction proceeds as follows. Let $x_1, \ldots, x_d$, where $d = VC(\Pi) - 1$, be a set shattered by Π with the largest possible cardinality. Let

$$\begin{gathered}X \in \{x_1, \ldots, x_d\}, \;\; P(X = x_j) = \tfrac{1}{d}; \\ T \in \{0, 1\}, \;\; P(T = 1) = p, \;\; T \perp (X, Y_0, Y_1); \\ Y_0 = 0,\end{gathered}$$

and, given a parameter vector $c = (c_1, \ldots, c_d) \in \{-1, 1\}^d$,

$$Y_1 | X = x_j = \begin{cases} A & \text{w.p. } \frac{1}{2}(1 + c_j\frac{\gamma}{A}) \\ -A & \text{w.p. } \frac{1}{2}(1 - c_j\frac{\gamma}{A}) \end{cases},$$

where $\gamma/A \leqslant 1$. Then, for $Y = TY_1 + (1-T)Y_0$,

$$\mathbb{E}(Y^2) = pA^2,$$
$$\tau(x_j) = \mathbb{E}[Y_1 - Y_0 | X = x_j] = \gamma c_j.$$

For every $c \in \{-1,1\}^d$, the joint distribution of $W = (Y, X, T)$ constructed above belongs to $P_{B,\eta}$ as long as $p \in [\eta, 1-\eta]$ and $pA^2 \leqslant B^2$. We will specify such p and A later.

Let $C = (C_1, \ldots, C_d)$ consist of i.i.d. random variables $C_j \in \{-1, 1\}$ such that $P(C_j = 1) = 1/2$. The joint distribution of $W = (Y, X, T)$ given $C = c$ is

$$P(Y = y, X = x_j, T = t | C = c) = \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}(1 + c_j \frac{\gamma}{A})\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}(1 - c_j \frac{\gamma}{A})\frac{p}{d} & y = -A, t = 1 \end{cases}.$$

We shall also derive the posterior probability $P(C_j = 1 | W_1^n)$ which will play a crucial role in deriving the lower bound.

We have

$$P(Y = y, X = x_j, T = t) = \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}\frac{p}{d} & y = -A, t = 1 \end{cases},$$

and

$$\begin{aligned} P(Y = y, X = x_k, T = t | C_j = 1) &= \mathbf{1}(k \neq j) P(Y = y, X = x_j, T = t) \\ &\quad + \mathbf{1}(k = j) \begin{cases} (1-p)\frac{1}{d} & y = 0, t = 0 \\ \frac{1}{2}(1 + \frac{\gamma}{A})\frac{p}{d} & y = A, t = 1 \\ \frac{1}{2}(1 - \frac{\gamma}{A})\frac{p}{d} & y = -A, t = 1 \end{cases}. \end{aligned}$$

Therefore,

$$\frac{P(W_i|C_j=1)}{P(W_i)} = \mathbf{1}(X_i \neq x_j) + \mathbf{1}(X_i = x_j) \begin{cases} 1 & Y_i = 0, T_i = 0 \\ 1 + \frac{\gamma}{A} & Y_i = A, T_i = 1 \\ 1 - \frac{\gamma}{A} & Y_i = -A, T_i = 1 \end{cases},$$

and

$$P(C_j = 1|W_1^n) = \frac{P(W_1^n|C_j=1)P(C_j=1)}{P(W_1^n)} = \frac{1}{2}\left(1 + \frac{\gamma}{A}\right)^{N_j^+}\left(1 - \frac{\gamma}{A}\right)^{N_j^-}, \tag{3.10}$$

where

$$N_j^+ = \#\{i : X_i = x_j, Y_i = A, T_i = 1\}$$
$$N_j^- = \#\{i : X_i = x_j, Y_i = -A, T_i = 1\},$$

so that a tuple $(N_j^+, N_j^-, n - N_j^+ - N_j^-)$ has a multinomial distribution:

$$P(N_j^+ = k_1, N_j^- = k_2|C_j = 1)$$
$$= \binom{n}{k_1}\binom{n-k_1}{k_2}\left(\frac{1}{2}(1+\frac{\gamma}{B})\frac{p}{d}\right)^{k_1}\left(\frac{1}{2}(1-\frac{\gamma}{B})\frac{p}{d}\right)^{k_2}\left(1-\frac{p}{d}\right)^{n-k_1-k_2}. \tag{3.11}$$

Now we turn to the main part of the proof. Let $\mathcal{P}_\mathcal{C} = \{P_{W|C=c} : c \in \{-1,1\}^d\} \subset \mathcal{P}_{B,\eta}$ denote the set of distributions of $W = (Y, X, T)$ constructed above, and μ denote the distribution of C. Let π_P^* denote the first-best treatment rule when the distribution of the data is P, and write $\pi_c^* = \pi^*_{P_{W|C=c}}$ for brevity. By construction, $\pi_c^*(x_j) = \mathbf{1}(c_j = 1)$, and $\pi_c^* \in \Pi$ since the class Π shatters $\{x_1, \ldots, x_d\}$. Note that:

$$V(\pi_c^*) - V(\hat{\pi}_n) = \frac{\gamma}{d}\sum_{j=1}^d c_j(\pi_c^*(x_j) - \hat{\pi}_n(x_j)) = \frac{\gamma}{d}\sum_{j=1}^d \mathbf{1}(\pi_c^*(x_j) \neq \hat{\pi}_n(x_j)).$$

Then,

$$
\begin{aligned}
\sup_{P\in\mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] &\geqslant \max_{P\in\mathcal{P}_{\mathcal{C}}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n))] \\
&\geqslant \int \mathbb{E}_{P_{W_1^n|C=c}}[V(\pi_c^*) - V(\hat{\pi}_n))] d\mu(c) \\
&= \frac{\gamma}{d} \sum_{j=1}^{d} \int\int \mathbf{1}(\pi_c^*(x_j) \neq \hat{\pi}_n(x_j)) dP_{W_1^n|C=c} d\mu(c) \qquad (3.12) \\
&= \frac{\gamma}{d} \sum_{j=1}^{d} P_{W_1^n,C_j}(\mathbf{1}(C_j = 1) \neq \hat{\pi}_n(x_j)) \\
&\geqslant \gamma \cdot \inf_{\pi} P_{W_1^n,C_j}(\mathbf{1}(C_j = 1) \neq \pi(W_1^n)).
\end{aligned}
$$

Note that $P_{W_1^n,C_j}(\mathbf{1}(C_j = 1) \neq \pi(W_1^n))$ is the probability of misclassification of $\mathbf{1}(C_j = 1)$ using W_1^n. By Theorem 2.1. in Devroye and Lugosi (1996), the infimum is attained by the Bayes Classifier, $\pi^*(W_1^n) = \mathbf{1}(P(C_j = 1|W_1^n) > 0.5)$, and is equal to

$$
\begin{aligned}
P(\mathbf{1}(C_j = 1) \neq \pi^*(W_1^n)) &= \tfrac{1}{2} P(\ P(C_j = 1|W_1^n) \leqslant 0.5 \ |C_j = 1) \\
&\quad + \tfrac{1}{2} P(\ P(C_j = 1|W_1^n) > 0.5 \ |C_j = -1).
\end{aligned}
$$

Denote $a = \gamma/A$, and work conditional on $C_j = 1$ from now on. Recalling (3.10),

$$
\begin{aligned}
P(P(C_j = 1|W_1^n) \leqslant 0.5) &= P((1+a)^{N_j^+}(1-a)^{N_j^-} \leqslant 1) \\
&\geqslant P((1-a^2)^{N_j^+} \leqslant 1 | N_j^+ \leqslant N_j^-) \cdot P(N_j^+ \leqslant N_j^-) \\
&= P(N_j^+ \leqslant N_j^-).
\end{aligned}
$$

Let $D_i^+ = \mathbf{1}(X_i = x_j, Y_i = A, T_i = 1)$ and $D_i^- = \mathbf{1}(X_i = x_j, Y_i = -A, T_i = 1)$. Then, $\mathbb{E}[D_i^+ - D_i^-] = ap/d$, $\mathbb{V}ar[D_i^+ - D_i^-] = p/d - (ap/d)^2$, and $\mathbb{E}[(D_i^+ - D_i^-)^3] = p/d$. Letting Z_n denote the studentized version of $n^{-1}\sum_{i=1}^n (D_i^+ - D_i^-)$ and Φ denote the Standard Normal CDF, using Berry-Esseen inequality we obtain

$$
\begin{aligned}
P(N_j^+ \leqslant N_j^-) &= P(\tfrac{1}{n}\textstyle\sum_{i=1}^n (D_i^+ - D_i^-) \leqslant 0) \\
&= P\left(Z_n \leqslant \frac{-\sqrt{n}ap/d}{\sqrt{p/d - (ap/d)^2}}\right) \\
&\geqslant \Phi\left(\frac{-\sqrt{n}ap/d}{\sqrt{p/d - (ap/d)^2}}\right) - \frac{K}{\sqrt{n}} \frac{1}{(p/d)^{1/2}(1-a^2p/d)^{3/2}},
\end{aligned}
$$

where $K < 0.469$ (Shevtsova, 2013). Choosing $a = \gamma/A \equiv c/\sqrt{n}\sqrt{d/p}$ for some $c \in (0,1)$, assuming n is large enough to satisfy $\gamma/A \leqslant 1$, we obtain

$$P(N_j^+ \leqslant N_j^-) \geqslant \Phi\left(-\frac{c}{\sqrt{1-c^2/n}}\right) - \frac{K}{\sqrt{n}}\frac{1}{\sqrt{p/d}(1-c^2/n)^{3/2}}.$$

Choosing $p = \eta$, $A = B/\sqrt{\eta}$ so that $\gamma = c \cdot B/\eta\sqrt{d/n}$, we have, for $n \geqslant 3$,

$$\begin{aligned}\sup_{P\in\mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] &\geqslant \tfrac{\gamma}{2} \cdot P(N_j^+ \leqslant N_j^- | C_j = 1) \\ &\geqslant \tfrac{1}{2}\tfrac{B}{\eta}\sqrt{\tfrac{d}{n}} \cdot c \cdot \Phi\left(-\tfrac{c}{\sqrt{1-c^2}}\right) - \tfrac{K}{2\sqrt{\eta}} \cdot \tfrac{B}{\eta}\tfrac{d}{n}\tfrac{c}{(1-c^2/3)^{3/2}}\end{aligned}$$

Choosing $c = 0.5162$, and plugging in $K = 0.469$ gives the final result

$$\sup_{P\in\mathcal{P}_{B,\eta}} \mathbb{E}_P[V(\pi_P^*) - V(\hat{\pi}_n)] \geqslant 0.07 \cdot \frac{B}{\eta}\sqrt{\frac{d}{n}} - \frac{0.14}{\sqrt{\eta}} \cdot \frac{B}{\eta}\frac{d}{n}.$$

For $n \geqslant 4d/\eta$, the right-hand-side in the preceding display is positive, and $\gamma/A \leqslant 1$ is also satisfied.

3.7.4 Proofs of Theorems 3.4 and 3.5

The proof of Theorem 3.4 is based on the following two lemmas. The first Lemma gives a maximal inequality in terms of the VC-dimension of the class of policy rules Π, the number of observations n, and the second moment of the orthogonal score Γ.

Lemma 3.8 (Uniform Concentration Bound for $\tilde{V}_n$)**.** *Suppose that the class Π has VC-dimension $VC(\Pi)$ and includes the no-treatment policy $\pi_0(x) = 0$ for all x. Then,*

$$\mathbb{E}\left[\sup_{\pi\in\Pi} |\tilde{V}_n(\pi) - V(\pi)|\right] \leqslant C\sqrt{\frac{VC(\Pi)S^2}{n}},$$

where $C \leqslant 58$ is a universal constant and $S^2 = \mathbb{E}(\Gamma_i^2)$.

Proof. Define a class of functions $\mathcal{F} = \{f(w) = \pi(x)\Gamma(w) : \pi \in \Pi\}$, which is a VC-subgraph class with $VC(\mathcal{F}) \leqslant VC(\Pi)$ and envelope $|\Gamma|$. Then, by Lemmas 3.1, 3.5, and the second

part of Lemma 3.7,

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|\tilde{V}_n(\pi)-V(\pi)|\right]\leqslant 2\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\pi(X_i)\Gamma_i\right|\right]\leqslant 2C\sqrt{\frac{VC(\Pi)S^2}{n}}.$$

where $C\leqslant 29$ is a constant from Lemma 3.7. ■

The second Lemma establishes that $\hat{V}_n$ and $\tilde{V}_n$ are uniformly close in $\pi\in\Pi$. It is a finite-sample version of Lemma 4 from Athey and Wager (2021) proven under slightly weaker assumptions.

Lemma 3.9 (Uniform Coupling)**.** *Let assumptions 3.2.1 – 3.2.4 hold, and assume that* $\mathbb{E}((Y-m(X,D))^2|X,D)\leqslant B^2$ *almost surely. Suppose that* $\hat{\Gamma}_i$ *are computed using a* J*-fold cross-fitting. Then,*

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|\hat{V}_n(\pi)-\tilde{V}_n(\pi)|\right]\leqslant R_{1,n}+R_{2,n}+R_{3,n},$$

where $C\leqslant 58$ *is a universal constant, and*

$$R_{1,n}=C\sqrt{(J+2)\cdot B^2\cdot\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_g}}}$$

$$R_{2,n}=C\sqrt{(J+2)\cdot\frac{2(\eta^2+1)}{\eta^2}\cdot\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}},$$

and

$$R_{3,n}=\sqrt{\frac{a((1-J^{-1})n)^2}{n^{\zeta_m+\zeta_g}}}.$$

Proof. Let $\hat{m}^{(-j)}$, $\tau_{\hat{m}^{(-j)}}$ and $\hat{g}^{(-j)}$ denote the estimators computed on observations excluding j-th fold. Denote the indices of the observations included in j-th fold by I_j. For an observation $i\in I_j$, write the difference $\hat{\Gamma}_i-\Gamma_i$ as a sum of three terms

$$\begin{aligned}\hat{\Gamma}_i-\Gamma_i &= (Y_i-m(X_i,T_i))(\hat{g}^{(-j)}(X_i,T_i)-g(X_i,T_i))\\ &+\tau_{\hat{m}^{(-j)}}(X_i,T_i)-\tau_m(X_i,T_i)-g(X_i,Z_i)(\hat{m}^{(-j)}(X_i,T_i)-m(X_i,T_i))\\ &-(\hat{g}^{(-j)}(X_i,Z_i)-g(X_i,Z_i))(\hat{m}^{(-j)}(X_i,T_i)-m(X_i,T_i))\end{aligned}$$

and denote the corresponding summands in $\hat{V}_n(\pi) - \tilde{V}_n(\pi)$ by $D_1(\pi), D_2(\pi)$, and $D_3(\pi)$. We will bound each term separately.

First Term. Write $D_1(\pi) = \sum_{j=1}^{J} D_1^{(j)}(\pi)$, where $\frac{n}{n_k} D_1^{(j)}(\pi)$ is equal to

$$\frac{1}{n_j} \sum_{i \in I_j} \pi(X_i)(Y_i - m(X_i, T_i))(\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i)).$$

Note that, by the law of iterated expectations,

$$\mathbb{E}[\pi(X_i)(Y_i - m(X_i, T_i))(\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i)) \mid \hat{g}^{(-j)}] = 0,$$

and denote the conditional second moment by

$$V_{1,n}(j) = \mathbb{E}\left[\pi(X_i)^2 \cdot \mathbb{E}[(Y_i - m(X_i, T_i))^2 | X_i, T_i] \cdot (\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i))^2 \mid g^{(-j)}\right].$$

Applying, conditional on $\hat{g}^{(-j)}$, Lemma 3.8 with $(Y_i - m(X_i, T_i)) \cdot (\hat{g}^{(-j)}(X_i, Z_i) - g(X_i, Z_i))$ in place of Γ_i, we get:

$$\frac{n}{n_j} \mathbb{E}\left[\sup_{\pi \in \Pi} |D_1^{(j)}(\pi)| \;\middle|\; \hat{g}^{(-j)}\right] \leqslant 2C \sqrt{\frac{VC(\Pi) V_{1,n}(j)}{n_j}}$$

Using Assumption 3.2.2, $\pi(X_i)^2 \leqslant 1$, and the bound on the conditional variance of Y,

$$\mathbb{E}(V_{1,n}(j)) \leqslant B \frac{a((\frac{J-1}{J})n)}{n^{\zeta_g}}$$

By the last two displays, the law of iterated expectations, and Jensen's inequality,

$$\mathbb{E}\left[\sup_{\pi \in \Pi} |D_1^{(j)}(\pi)|\right] \leqslant 2C \sqrt{\frac{n_j}{n}} \sqrt{B \frac{VC(\Pi) a((1 - J^{-1})n)}{n^{1+\zeta_g}}}$$

Since $n_j/n \leqslant 1/(J-1)$ and supremum is sub-additive,

$$\mathbb{E}\left[\sup_{\pi \in \Pi} |D_1(\pi)|\right] \leqslant 2C \sqrt{(J+2)B \cdot \frac{VC(\Pi) a((1 - J^{-1})n)}{n^{1+\zeta_g}}}$$

Second Term. As before, write $D_2(\pi) = \sum_{j=1}^{J} D_2^{(j)}(\pi)$, where $\frac{n}{n_j} D_2^{(j)}(\pi)$ is equal to

$$\frac{1}{n_j} \sum_{i \in I_j} \pi(X_i)(\tau_{\hat{m}^{(-j)}}(X_i, T_i) - \tau_m(X_i, T_i) - g(X_i, Z_i)(\hat{m}^{(-j)}(X_i, T_i) - m(X_i, T_i)))$$

Denote the individual summands in the previous display by $f(W_i;\pi)$. Note that

$$\mathbb{E}(f(W_i;\pi)|\hat{m}^{(-j)},\tau_{\hat{m}^{(-j)}})=0$$

by part (2) of Assumption 3.2.1 and the law of iterated expectations. Denote $V_{2,n}(j)=\mathbb{E}(f(W_i;\pi)^2|\hat{m}^{(-j)},\tau_{\hat{m}^{(-j)}})$. Applying, conditional on $\hat{m}^{(-j)}$ and $\tau_{\hat{m}^{(-j)}}$, Lemma 3.8 with $(\tau_{\hat{m}^{(-j)}}(X_i,T_i)-\tau_m(X_i,T_i)-g(X_i,Z_i)(\hat{m}^{(-j)}(X_i,T_i)-m(X_i,T_i)))$ in place of Γ_i, we get:

$$\frac{n}{n_j}\mathbb{E}\left[\sup_{\pi\in\Pi}|D_2^{(j)}(\pi)|\;\middle|\;\hat{g}^{(-j)}\right]\leqslant 2C\sqrt{\frac{VC(\Pi)V_{2,n}(j)}{n_j}}$$

Using $(a+b)^2\leqslant 2(a^2+b^2)$, $\pi(X_i)^2\leqslant 1$, and Assumptions 3.2.2 and 3.2.4, we get:

$$\mathbb{E}(V_{2,n}(j))\leqslant 2\left(\frac{a((1-J^{-1})n)}{n^{\zeta_m}}+\frac{1}{\eta^2}\frac{a((1-J^{-1})n)}{n^{\zeta_m}}\right)=\frac{2(\eta^2+1)}{\eta^2}\frac{a((1-J^{-1})n)}{n^{\zeta_m}}.$$

By the last two displays, the law of iterated expectation, and Jensen's inequality:

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|D_2^{(j)}(\pi)|\right]\leqslant 2C\sqrt{\frac{n_j}{n}}\sqrt{\frac{2(\eta^2+1)}{\eta^2}\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}}$$

Since $n_j/n\leqslant 1/(J-1)$ and supremum is sub-additive,

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|D_1(\pi)|\right]\leqslant 2C\sqrt{(J+2)\frac{2(\eta^2+1)}{\eta^2}\frac{VC(\Pi)a((1-J^{-1})n)}{n^{1+\zeta_m}}}$$

Third Term. Let $j(i)$ denote the fold in which observation i belongs. We have:

$$D_3(\pi)=-\frac{1}{n}\sum_{i=1}^n\pi(X_i)(\hat{g}^{(-j(i))}(X_i,Z_i)-g(X_i,Z_i))(\hat{m}^{(-j(i))}(X_i,T_i)-m(X_i,T_i))$$

By Cauchy-Schwartz inequality and $\pi(X_i)^2\leqslant 1$,

$$\begin{aligned}|D_3(\pi)|\;&\leqslant\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n(\hat{g}^{(-j(i))}(X_i,Z_i)-g(X_i,Z_i))^2}\\&\quad\times\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n(\hat{m}^{(-j(i))}(X_i,T_i)-m(X_i,T_i))^2},\end{aligned}$$

where we note that the right hand side does not depend on π. Taking expectations on both sides, using Cauchy-Schwartz inequality one more time, and recalling Assumption 3.2.2, we obtain

$$\mathbb{E}\left[\sup_{\pi\in\Pi}|D_3(\pi)|\right]\leqslant\sqrt{\frac{a((1-J^{-1})n)^2}{n^{\zeta_m+\zeta_g}}},$$

and the proof is complete. ■

3.7.4.1 Proof of Theorem 3.4

To keep the notation simple, we write $\hat{\pi}_n$ instead of $\hat{\pi}_n^{REWM}$ and write $\mathbb{E}$ instead of $\mathbb{E}_P$ for a fixed distribution $P \in \mathcal{P}_{B_\tau,B,\eta}$. Let $\pi^* \in \Pi$ be such that $V(\pi^*) = \max_{\pi\in\Pi} V(\pi)$. Note that:

$$\begin{aligned} R(\hat{\pi}_n) &= V(\pi^*) - V(\hat{\pi}_n) \\ &= V(\pi^*) - \hat{V}_n(\hat{\pi}_n) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n) \\ &\leqslant V(\pi^*) - \hat{V}_n(\pi^*) + \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n). \end{aligned}$$

Then, writing

$$\begin{aligned} V(\pi^*) - \hat{V}_n(\pi^*) &= V(\pi^*) - \tilde{V}_n(\pi^*) + \tilde{V}_n(\pi^*) - \hat{V}_n(\pi^*) \\ \hat{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n) &= \hat{V}_n(\hat{\pi}_n) - \tilde{V}_n(\hat{\pi}_n) + \tilde{V}_n(\hat{\pi}_n) - V(\hat{\pi}_n), \end{aligned}$$

and using $\mathbb{E}[V(\pi^*) - \tilde{V}(\pi^*)] = 0$, we obtain:

$$\mathbb{E}[R(\hat{\pi}_n)] \leqslant \mathbb{E}[\sup_{\pi\in\Pi} |\tilde{V}_n(\pi) - V(\pi)|] + 2\mathbb{E}[\sup_{\pi\in\Pi} |\hat{V}_n(\pi) - \tilde{V}_n(\pi)|]. \tag{3.13}$$

By Lemma 3.8, the first term is bounded by $2C\sqrt{VC(\Pi)S^2/n}$, where $S^2 = \mathbb{E}[\Gamma^2]$. By the Law of Iterated Expectations and $P \in \mathcal{P}_{B_\tau,B,\eta}$,

$$\begin{aligned} \mathbb{E}[\Gamma^2] &= \mathbb{E}[(\tau_m(X,T) + g(X,Z)(Y - m(X,T)))^2] \\ &= \mathbb{E}[\tau_m^2(X,T)] + \mathbb{E}[g(X,Z)^2(Y - m(X,T))^2] \\ &\leqslant B_\tau^2 + \eta^{-2}B^2. \end{aligned}$$

The second term in (3.13) is bounded by Lemma 3.9, so the desired result follows.

Before proving the main result of the paper, we include another technical lemma for easier reference.

Lemma 3.10 (Addendum to Lemma 3.9)**.** *Let W_1^l denote the estimating sample with $l = (1-s)n$. In the notation of Lemma 3.9:*

1. *For every fixed $\pi \in \Pi$:*

$$\mathbb{E}[\hat{V}_l(\pi) - \tilde{V}_l(\pi)] \leqslant R_{3,l}.$$

2. For any $\hat{\pi}_{l,k}$ computed using the estimated sample W_1^l,

$$\mathbb{E}[\hat{V}_r(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k})] \leqslant R_{3,l}.$$

Proof. To prove the first claim, we apply the same argument as in Lemma 3.9. The expectations of the first two corresponding terms, denoted there by $D_1(\pi)$ and $D_2(\pi)$, are equal to zero, and the expectation of the third term is shown to be less than $R_{3,l}$.

The proof of the second claim is easier, since we do not need to separate the contributions of different folds. Replacing the arguments of the functions with the index of the observation (from the test sample) to which they are applied, we can expand $\hat{\Gamma}_i - \Gamma_i$ as a sum of three terms:

$$\hat{\Gamma}_i - \Gamma_i = (\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i)) + (Y_i - m_i)(\hat{g}_i - g_i) - (\hat{m}_i - m_i)(\hat{g}_i - g_i).$$

Let D_1, D_2 and D_3 denote the corresponding terms in $\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})$. Then, by Assumption 3.2.1-2 and the Law of Iterated Expectations,

$$\mathbb{E}[D_1|W_1^l] = \mathbb{E}\left[\hat{\pi}_{l,k}(X_i) \cdot \mathbb{E}[(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))|X_i, W_1^l] \,\middle|\, W_1^l\right] = 0.$$

Further, by the Law of Iterated Expectations and the exclusion restriction on Z_i,

$$\mathbb{E}[D_2|W_1^l] = \mathbb{E}\left[\hat{\pi}_{l,k}(X_i) \cdot \mathbb{E}[Y_i - m_i|X_i, T_i, W_1^l] \cdot (\hat{g}_i - g_i) \,\middle|\, W_1^l\right] = 0.$$

Finally, by Cauchy-Schwartz inequality and $\hat{\pi}_{l,k}(X_i)^2 \leqslant 1$,

$$D_3 \leqslant \sqrt{\tfrac{1}{r}\textstyle\sum_i (\hat{m}_i - m_i)^2} \cdot \sqrt{\tfrac{1}{r}\textstyle\sum_i (\hat{g}_i - g_i)^2}.$$

Taking expectations on both sides, applying Cauchy-Schwartz inequality again, and using the Law of Iterated Expectations, we obtain

$$\mathbb{E}[D_3] \leqslant \sqrt{\mathbb{E}[(\hat{m}_i - m_i)^2] \cdot \mathbb{E}[(\hat{g}_i - g_i)^2]} \leqslant R_{3,l},$$

■

3.7.4.2 Proof of Theorem 3.5

To keep the notation simple, we write $\hat{\pi}_{n,\hat{k}}$ instead of $\hat{\pi}_n^{RPWM}$ and $\mathbb{E}$ instead of $\mathbb{E}_P$ for a fixed distribution $P \in \mathcal{P}_{B_\tau,B,\eta}$. The subscripts l, r, and n indicate that the corresponding object depends only on the estimating sample, only on the test sample, or on the entire sample. For example, while $\hat{\pi}_{l,k}$ only depends on the estimating sample, $\hat{\pi}_{n,\hat{k}}$ depends on the entire sample due to the choice of $\hat{k}$. Let $\pi_k^* \in \Pi_k$ be such that $V(\pi_k^*) = V_{\Pi_k}^*$. Write:

$$V_\Pi^* - V(\hat{\pi}_{n,\hat{k}}) = V_\Pi^* - V_{\Pi_k}^* + \underbrace{V_{\Pi_k} - Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}})}_{(I)} + \underbrace{Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})}_{(II)}. \tag{3.14}$$

First, since $Q_{n,\hat{k}}(\hat{\pi}_{n,\hat{k}}) \geqslant Q_{n,k}(\hat{\pi}_{l,k})$, and $\hat{V}_l(\hat{\pi}_{l,k}) \geqslant \hat{V}_l(\pi_k^*)$, we can bound:

$$\begin{aligned}(I) &\leqslant V(\pi_k^*) - Q_{n,k}(\hat{\pi}_{l,k}) \\ &\leqslant V(\pi_k^*) - \hat{V}_l(\pi_k^*) + \hat{C}_{n,k} \\ &= V(\pi_k^*) - \tilde{V}_l(\pi_k^*) + \tilde{V}_l(\pi_k^*) - \hat{V}_l(\pi_k^*) + \hat{C}_{n,k}.\end{aligned}$$

Here, $\mathbb{E}[V(\pi_k^*) - \tilde{V}_l(\pi_k^*)] = 0$ and, by Lemma 3.10, $\mathbb{E}[\tilde{V}_l(\pi_k^*) - \hat{V}_l(\pi_k^*)] \leqslant R_{3,l}$. Therefore,

$$\mathbb{E}[(I)] \leqslant \mathbb{E}[\hat{C}_{n,k}] + R_{3,l}.$$

Next, consider

$$(II) = \hat{V}_r(\hat{\pi}_{n,\hat{k}}) - \tilde{V}_r(\hat{\pi}_{n,\hat{k}}) + (\tilde{V}_r(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})).$$

The first summand can be bounded by

$$\begin{aligned}\mathbb{E}\left[\hat{V}_r(\hat{\pi}_{n,\hat{k}}) - \tilde{V}_r(\hat{\pi}_{n,\hat{k}})\right] &\leqslant \mathbb{E}\left[\max_{k\leqslant K} |\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})|\right] \\ &\leqslant K \max_{k\leqslant K} \mathbb{E}\left[|\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})|\right].\end{aligned}$$

As in the proof of Lemma 3.9, we can expand:

$$\hat{\Gamma}_i - \Gamma_i = (\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i)) + (Y_i - m_i)(\hat{g}_i - g_i) - (\hat{m}_i - m_i)(\hat{g}_i - g_i),$$

so that:

$$\begin{aligned}\mathbb{E}\left[|\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})|\right] &= \mathbb{E}\left[|\tfrac{1}{r}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{\Gamma}_i - \Gamma_i)|\right] \\ &\leqslant \tfrac{1}{\sqrt{r}}\mathbb{E}\left[|\tfrac{1}{\sqrt{r}}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))|\right] \\ &+ \tfrac{1}{\sqrt{r}}\mathbb{E}\left[|\tfrac{1}{\sqrt{r}}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(Y_i - m_i)(\hat{g}_i - g_i)|\right] \\ &+ \mathbb{E}\left[|\tfrac{1}{r}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)|\right]\end{aligned}$$

By Assumption 3.2.1-2 and the Law of Iterated Expectations,

$$\mathbb{E}[\hat{\pi}_{l,k}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))|W_1^n, X_i] = 0.$$

Using $\mathbb{E}[|W|]^2 \leqslant \mathbb{E}[W^2]$, the Law of Iterated Expectations, $\hat{\pi}_{l,k}^2(X_i) \leqslant 1$, and Assumption 3.2.2, we obtain:

$$\begin{aligned}\mathbb{E}\left[|\tfrac{1}{\sqrt{r}}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))|\right]^2 &\leqslant \mathbb{E}\left[\tfrac{1}{r}\textstyle\sum_i(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))^2\right] \\ &= \mathbb{E}[(\tau_{\hat{m},i} - \tau_{m,i} - g_i(\hat{m}_i - m_i))^2] \\ &\leqslant 2(\mathbb{E}[(\tau_{\hat{m},i} - \tau_{m,i})^2] + \mathbb{E}[g_i^2(\hat{m}_i - m_i)^2]) \\ &\leqslant 2\tfrac{\eta^2+1}{\eta^2}\tfrac{a((1-J^{-1})l)}{l^{\zeta_m}}.\end{aligned}$$

A similar argument and the bound $\mathbb{E}[(Y_i - m_i)^2|X_i, T_i] \leqslant B^2$ yield:

$$\mathbb{E}\left[|\tfrac{1}{\sqrt{r}}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(Y_i - m_i)(\hat{g}_i - g_i)|\right]^2 \leqslant \mathbb{E}[(Y_i - m_i)^2(\hat{g}_i - g_i)^2] \leqslant B^2 \cdot \tfrac{a((1-J^{-1})l)}{l^{\zeta_g}}.$$

Next, by Cauchy-Schwartz inequality and $\hat{\pi}_{l,k}^2(X_i) \leqslant 1$,

$$|\tfrac{1}{r}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)| \leqslant \sqrt{\tfrac{1}{r}\textstyle\sum_i(\hat{m}_i - m_i)^2} \cdot \sqrt{\tfrac{1}{r}\textstyle\sum_i(\hat{g}_i - g_i)^2}$$

Taking expectations on both sides, applying Cauchy-Schwartz inequality and the Law of Iterated Expectations,

$$\mathbb{E}[|\tfrac{1}{r}\textstyle\sum_i \hat{\pi}_{l,k}(X_i)(\hat{m}_i - m_i)(\hat{g}_i - g_i)|] \leqslant \sqrt{\mathbb{E}[(\hat{m}_i - m_i)^2] \cdot \mathbb{E}[(\hat{g}_i - g_i)^2]} \leqslant \sqrt{\tfrac{a((1-J^{-1})l)}{l^{\zeta_m+\zeta_g}}}$$

Combining the above results, we obtain:

$$\mathbb{E}\left[|\hat{V}_r(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})|\right] \leqslant \sqrt{\tfrac{1}{s(1-s)^{\zeta_m \wedge \zeta_g}}}\sqrt{\tfrac{2(\eta^2+1)}{\eta^2} \vee B^2}\sqrt{\tfrac{a((1-J^{-1})(1-s)n)}{n^{1+\zeta_m \wedge \zeta_g}}} + R_{3,(1-s)n}$$

For the second summand in (II), arguing as in the proof of Theorem 3.3 (see Equations (3.7), (3.8), and the following argument and recall that $\hat{V}_n$ in that proof plays the same role as $\tilde{V}_r$ in this one),

$$\mathbb{E}\left[\tilde{V}_r(\hat{\pi}_{n,\hat{k}}) - V(\hat{\pi}_{n,\hat{k}})\right] \leqslant \sqrt{B_\tau^2 + \eta^{-2}B^2}\frac{K}{\sqrt{sn}} = K\frac{\sqrt{B_\tau^2\eta^2+B^2}}{\eta}\sqrt{\frac{1}{sn}}.$$

Let $R_{3,(1-s)n}$ denote the rate Lemma 3.9 with $(1-s)n$ instead of n. Defining

$$S_{2,n} \equiv \sqrt{\frac{1}{s(1-s)^{\zeta_m\wedge\zeta_g}}}\sqrt{\frac{2(\eta^2+1)}{\eta^2}\vee B^2}\sqrt{\frac{a((1-J^{-1})(1-s)n)}{n^{1+\zeta_m\wedge\zeta_g}}} + 2R_{3,(1-s)n} \tag{3.15}$$

and

$$S_n \equiv K\frac{\sqrt{B_\tau^2\eta^2+B^2}}{\eta}\sqrt{\frac{1}{sn}} + S_{2,n}, \tag{3.16}$$

we conclude that

$$\mathbb{E}[(I)+(II)] \leqslant \mathbb{E}[\hat{C}_{n,k}] + S_n.$$

Therefore, for any $k \leqslant K$,

$$\mathbb{E}[R(\hat{\pi}_{n,\hat{k}})] \leqslant V_\Pi^* - V_{\Pi_k}^* + \mathbb{E}[\hat{C}_{n,k}] + S_n, \tag{3.17}$$

and the first statement of the Theorem follows from taking an infimum over $k \leqslant K$.

To prove the second statement, it remains to bound $\mathbb{E}[\hat{C}_{n,k}]$. To this end, write:

$$\begin{aligned}\hat{C}_{n,k} &= \tilde{V}_l(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k}) + \hat{V}_l(\hat{\pi}_{l,k}) - \tilde{V}_l(\hat{\pi}_{l,k}) \\ &+ \tilde{V}_r(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k}) \\ &+ V(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k}).\end{aligned}$$

By Lemmas 3.8 and 3.9, for any $P \in \mathcal{P}_{B_\tau,B,\eta}$,

$$\begin{aligned}\mathbb{E}[\tilde{V}_l(\hat{\pi}_{l,k}) - V(\hat{\pi}_{l,k})] &\leqslant C\frac{\sqrt{B_\tau^2\eta^2+B^2}}{\eta}\sqrt{\frac{VC(\Pi_k)}{(1-s)n}}, \\ \mathbb{E}[\hat{V}_l(\hat{\pi}_{l,k}) - \tilde{V}_l(\hat{\pi}_{l,k})] &\leqslant R^k_{1,(1-s)n} + R^k_{2,(1-s)n} + R_{3,(1-s)n},\end{aligned}$$

where $R^k_{j,(1-s)n}$, for $j = 1,2,3$, are defined in Lemma 3.9 with Π_k instead of Π and $(1-s)n$ instead of n. Finally, by Lemma 3.10, $\mathbb{E}[\tilde{V}_r(\hat{\pi}_{l,k}) - \hat{V}_r(\hat{\pi}_{l,k})] \leqslant R_{3,(1-s)n}$, and by the Law of

Iterated Expectations, $\mathbb{E}[V(\hat{\pi}_{l,k}) - \tilde{V}_r(\hat{\pi}_{l,k})] = 0$. Plugging the above into (3.17), and noting that for every $P \in \mathcal{P}^k_{B_\tau,B,\eta}$, we have $V^*_\Pi = V^*_{\Pi_k}$,

$$\sup_{P \in \mathcal{P}^k_{B_\tau,B,\eta}} \mathbb{E}_P[R(\hat{\pi}_{n,\hat{k}})] \leqslant \frac{\sqrt{B_\tau^2\eta^2+B^2}}{\eta}\left(C\sqrt{\frac{VC(\Pi_k)}{(1-s)n}} + K\sqrt{\frac{1}{sn}}\right) + S^k_{1,n} + S_{2,n},$$

where $S^k_{1,n} = R^k_{1,(1-s)n} + R^k_{2,(1-s)n} + R_{3,(1-s)n}$, and $S_{2,n}$ is given in Equation 3.15.

Bibliography

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, 113(2):231–263.

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014). Asymptotic efficiency of semiparametric two-step gmm. *Review of Economic Studies*, 81(3):919–943.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2):442–457.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

Aradillas-López, A., Gandhi, A., and Quint, D. (2013a). Identification and inference in ascending auctions with correlated private values. *Econometrica*, 81(2):489–534.

Aradillas-López, A., Gandhi, A., and Quint, D. (2013b). Identification and inference in ascending auctions with correlated private values. *Econometrica*, 81(2):489–534.

Armstrong, T. and Shen, S. (2015). Inference on optimal treatment assignments.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (2008). *A first course in order statistics*. SIAM.

Artstein, Z. (1983). Distributions of random sets and random selections. *Israel Journal of Mathematics*, 46(4):313–324.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Barseghyan, L., Coughlin, M., Molinari, F., and Teitelbaum, J. C. (2021). Heterogeneous choice sets and preferences. *Econometrica*, 89(5):2015–2048.

Bassett, G. and Koenker, R. (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, 77(378):407–415.

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

Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, 76(4):763–814.

Bhattacharya, D. and Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Blumenthal, S. and Cohen, A. (1968). Estimation of the larger translation parameter. *The Annals of Mathematical Statistics*, pages 502–516.

Blundell, R., Gosling, A., Ichimura, H., and Meghir, C. (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, 75(2):323–363.

Bogachev, V. I. (1998). *Gaussian measures*. Number 62. American Mathematical Soc.

Bogachev, V. I. (2007). *Measure theory*, volume 1. Springer Science & Business Media.

Bontemps, C., Magnac, T., and Maurin, E. (2012). Set identified linear models. *Econometrica*, 80(3):1129–1155.

Brown, B. W. and Newey, W. K. (1998). Efficient semiparametric estimation of expectations. *Econometrica*, 66(2):453–464.

Bugni, F. A. (2016). Comparison of inferential methods in partially identified models in terms of error in coverage probability. *Econometric Theory*, 32(1):187–242.

Cai, T. T. and Low, M. G. (2011). Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041.

Canay, I. A. and Shaikh, A. M. (2017). Practical and theoretical advances in inference for partially identified models. *Advances in Economics and Econometrics*, 2:271–306.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics*, 34(3):305–334.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 567–596.

Chen, X. and Santos, A. (2018). Overidentification in regular models. *Econometrica*, 86(5):1771–1817.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2016). Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013). Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737.

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

Chesher, A. and Rosen, A. M. (2020). Generalized instrumental variable models, methods, and applications. In *Handbook of Econometrics*, volume 7, pages 1–110. Elsevier.

Chesher, A., Rosen, A. M., and Smolinski, K. (2013). An instrumental variable model of multiple discrete choice. *Quantitative Economics*, 4(2):157–196.

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

De Paula, Á., Richards-Shubik, S., and Tamer, E. (2018). Identifying preferences in networks with bounded degree. *Econometrica*, 86(1):263–288.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, 125(1-2):141–173.

Doss, H. and Sethuraman, J. (1989). The price of bias reduction when there is no unbiased estimate. *The Annals of Statistics*, pages 440–442.

Dvoretzky, A., Wald, A., and Wolfowitz, J. (1951). Elimination of randomization in certain statistical decision procedures and zero-sum two-person games. *The Annals of Mathematical Statistics*, pages 1–21.

Fang, Z. (2018). Optimal plug-in estimators of directionally differentiable functionals. *Working paper*.

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

Feinberg, E. A. and Piunovskiy, A. B. (2006). On the dvoretzky–wald–wolfowitz theorem on nonrandomized statistical decisions. *Theory of Probability & Its Applications*, 50(3):463–466.

Galichon, A. and Henry, M. (2011). Set identification in models with multiple equilibria. *The Review of Economic Studies*, 78(4):1264–1298.

Gualdani, C. (2021). An econometric model of network formation with an application to board interlocks between firms. *Journal of Econometrics*, 224(2):345–370.

Haile, P. A. and Tamer, E. (2003). Inference with an incomplete model of english auctions. *Journal of Political Economy*, 111(1):1–51.

Hansen, B. E. (2017). Regression kink with an unknown threshold. *Journal of Business & Economic Statistics*, 35(2):228–240.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701.

Hirano, K. and Porter, J. R. (2012). Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790.

Ho, K. and Rosen, A. M. (2015). Partial identification in applied research: benefits and challenges. Technical report, National Bureau of Economic Research.

Hong, H. and Li, J. (2020). The numerical bootstrap. *The Annals of Statistics*, 48(1):397–412.

Horowitz, J. L. (2001). The bootstrap. In *Handbook of econometrics*, volume 5, pages 3159–3228. Elsevier.

Ibragimov, I. A. and Hasḿinskii, R. Z. (1981). *Statistical estimation: asymptotic theory*. Springer, New York, NY.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Kaido, H. and Santos, A. (2014). Asymptotically efficient estimation of models defined by convex moment inequalities. *Econometrica*, 82(1):387–413.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

Kline, P. and Santos, A. (2013). Sensitivity to missing data assumptions: Theory and evaluation of the us wage structure. *Quantitative Economics*, 4(2):231–267.

Koshevnik, Y. A. and Levit, B. Y. (1976). On a non-parametric analogue of the information matrix. *Teoriya Veroyatnostei i ee Primeneniya*, 21(4):759–774.

Kreider, B. and Pepper, J. V. (2007). Disability and employment: Reevaluating the evidence in light of reporting errors. *Journal of the American Statistical Association*, 102(478):432–441.

Kreider, B., Pepper, J. V., Gundersen, C., and Jolliffe, D. (2012). Identifying the effects of snap (food stamps) on child health outcomes when participation is endogenous and misreported. *Journal of the American Statistical Association*, 107(499):958–975.

Le Cam, L. M. (1986). *Asymptotic methods in statistical decision theory*. Springer Verlag.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Luo, Y. and Wang, H. (2018). Identifying and computing the exact core-determining class. *Available at SSRN 3154285*.

Manski, C. F. (2003). *Partial identification of probability distributions*, volume 5. Springer.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

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

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1010.

Manski, C. F. and Pepper, J. V. (2009). More on monotone instrumental variables. *The Econometrics Journal*, 12:S200–S216.

Manski, C. F. and Sims, C. (1994). The selection problem. In *Advances in Econometrics, Sixth World Congress*, volume 1, pages 143–70.

Manski, C. F. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70(2):519–546.

Masten, M. A. and Poirier, A. (2020). Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111.

Mbakop, E. and Tabord-Meehan, M. (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89(2):825–848.

Molchanov, I. and Molinari, F. (2018). *Random sets in econometrics*, volume 60. Cambridge University Press.

Molinari, F. (2020). Microeconometrics with partial identification. *Handbook of econometrics*, 7:355–486.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research*, 6(1):58–73.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135.

Newey, W. K. (1994a). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Newey, W. K. (1994b). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51.

Pakes, A. (2010). Alternative models for moment inequalities. *Econometrica*, 78(6):1783–1822.

Pakes, A., Porter, J., Ho, K., and Ishii, J. (2007). Moment inequalities and their application. Technical report, CEMMAP Working paper.

Pakes, A., Porter, J., Ho, K., and Ishii, J. (2015). Moment inequalities and their application. *Econometrica*, 83(1):315–334.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. Springer, New York, NY.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pages 1403–1430.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2014). A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Shaikh, A. M. and Vytlacil, E. J. (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica*, 79(3):949–955.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3):477–487.

Sheng, S. (2020). A structural econometric analysis of network formation games through subnetworks. *Econometrica*, 88(5):1829–1858.

Song, K. (2014). Local asymptotic minimax estimation of nonregular parameters with translation-scale equivariant maps. *Journal of Multivariate Analysis*, 125:136–158.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81.

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

Tebaldi, P., Torgovitsky, A., and Yang, H. (2019). Nonparametric estimates of demand in the california health insurance exchange. Technical report, National Bureau of Economic Research.

Telgen, J. (1983). Identifying redundant constraints and implicit equalities in systems of linear constraints. *Management Science*, 29(10):1209–1222.

Tibshirani, R. J. and Efron, B. (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436.

van der Vaart, A. W. (1988). Statistical estimation in large parameter spaces. *CWI Tracts*.

van der Vaart, A. W. (1991). On differentiable functionals. *The Annals of Statistics*, pages 178–204.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence*. Springer.