Optimal Plug-in Estimators of Directionally Differentiable Functionals

(Job Market Paper)

Zheng Fang*
Department of Economics, UC San Diego
zfang@ucsd.edu

November, 2014

Abstract

This paper studies optimal estimation of parameters taking the form $\phi(\theta_0)$, where θ_0 is unknown but can be regularly estimated and ϕ is a known directionally differentiable function. The irregularity caused by nondifferentiability of ϕ makes traditional optimality criteria such as semiparametric efficiency and minimum variance unbiased estimation impossible to apply. We instead consider optimality in the sense of local asymptotic minimaxity – i.e. we seek estimators that locally asymptotically minimize the maximum of the risk function. We derive the lower bound of local asymptotic minimax risk within a class of plug-in estimators and develop a general procedure for constructing estimators that attain the bound. As an illustration, we apply the developed theory to the estimation of the effect of Vietnam veteran status on the quantiles of civilian earnings.

KEYWORDS: Directional differentiability, Local asymptotic minimaxity, Plug-in

*I am deeply grateful to my advisor, Andres Santos, for his constant support and guidance. I would also like to thank Brendan Beare and Graham Elliott for their comments and suggestions that helped greatly improve this paper.

1 Introduction

In many econometric problems, parameters of interest embody certain irregularity that presents significant challenges for estimation and inference (Hirano and Porter, 2012; Fang and Santos, 2014). A large class of these parameters take the form $\phi(\theta_0)$ where θ_0 is a well-behaved parameter that depends on the underlying distribution of the data while ϕ is a known but potentially nondifferentiable function. Economic settings in which such irregularity arises with ease include treatment effects (Manski and Pepper, 2000; Hirano and Porter, 2012; Song, 2014a,b), interval valued data (Manski and Tamer, 2002), incomplete auction models (Haile and Tamer, 2003), and estimation under shape restrictions (Chernozhukov et al., 2010).

The aforementioned examples share the common feature of ϕ being directionally differentiable despite a possible failure of full differentiability. In this paper, we study optimal estimation of $\phi(\theta_0)$ for such irregular ϕ. In regular settings, one usually thinks of optimality in terms of semiparametric efficiency (Bickel et al., 1993). Unfortunately, the irregularity caused by nondifferentiability of ϕ makes traditional optimality criteria including semiparametric efficiency impossible to apply – in particular, if ϕ is nondifferentiable, then any estimator for $\phi(\theta_0)$ is necessarily irregular and biased(Hirano and Porter, 2012). Hence, the first question we need to address is: what is an appropriate notion of optimality for nondifferentiable ϕ? Following the decision theoretic framework initiated by Wald (1950) and further developed by Le Cam (1955, 1964), we may compare the competing estimators under consideration by examining their expected losses. Specifically, let T_n be an estimator of $\phi(\theta_0)$ and ℓ a loss function that measures the loss of estimating $\phi(\theta_0)$ using T_n by $\ell(r_n\{T_n - \phi(\theta_0)\})$, where $r_n \uparrow \infty$ is the rate of convergence for estimation of θ_0. The resulting expected loss or *risk function* is then

$$E_P[\ell(r_n\{T_n - \phi(\theta(P))\})] \ , \tag{1.1}$$

where E_P denotes the expectation taken with respect to P that generates the data and $\theta_0 \equiv \theta(P)$ signifies the dependence of θ_0 on P. The function (1.1) can in turn be employed to assess the performance of the estimator T_n – in particular, we would like an estimator to have the smallest possible risk at every P in the model. Unfortunately, it is well known that there exist no estimators that minimize the risk uniformly for all P (Lehmann and Casella, 1998).

As ways out of this predicament, one can either restrict the class of competing estimators, or seek an estimator that has the smallest risk in some overall sense. For the former approach, common restrictions imposed on estimators include mean unbiasedness, quantile unbiasedness and equivariance (including regularity which is also known as asymptotic equivariance in law). By Hirano and Porter (2012), however, if ϕ is only directionally differentiable, then no mean unbiased, quantile unbiased or regular estimators exist. It is noteworthy that non-existence of unbiased estimators implies that bias correction procedures cannot fully eliminate the bias of any estimator; in fact, any procedure that tries to remove the bias would push the variance to infinity (Doss and Sethuraman, 1989). As to equivariance in terms of groups of transformations, it is unclear to us what a suitable group of invariant transformations should be. Alternatively, one may translate the risk function (1.1) into a single number such as Bayesian risk that leads to average risk optimality or the maximum risk that leads to minimaxity. Since

our analysis shall focus on local risk, one may not have natural priors on the space of localization parameters in order to evaluate the Bayesian risk. Moreover, when the model is semiparametric or nonparametric which our setup accommodates, Bayes estimators entail specification of priors on infinite dimensional spaces which practitioners may lack.

The approach we adopt in this paper towards optimal estimation of $\phi(\theta_0)$ is a combination of the above two: we confine our attention to the important class of plug-in estimators of the form $\phi(\hat{\theta}_n)$, where $\hat{\theta}_n$ is a regular estimator of θ_0, and seek estimators that minimize the maximum of the risk function – i.e. the risk under the worst case scenario. In addition, the optimality shall be in local sense, that is, we consider maximum risk over neighborhoods around the distribution that generates the data. This is justified by the facts that global risk is somewhat too restrictive for infinite dimensional P (Bickel et al., 1993, p.21) and that one can locate the unknown parameter with considerable precision as sample size increases (Hájek, 1972). Specifically, for $\hat{\theta}_n$ an arbitrary regular estimator of θ_0, we establish the lower bound of the following local asymptotic minimax risk:

$$\sup_{I \subset_f H} \liminf_{n\to\infty} \sup_{h \in I} E_{P_{n,h}}[\ell(r_n\{\phi(\hat{\theta}_n) - \phi(\theta(P_{n,h}))\})] \ , \tag{1.2}$$

where H is the set of localization parameters, and $I \subset_f H$ signifies that I is a finite subset of H so that the first supremum is taken over all finite subsets of H.[1] For detailed explanations on why we take the above version of local asymptotic minimaxity, which dates back to van der Vaart (1988a, 1989), we defer our discussion to Section 2.3. The lower bound derived relative to the local asymptotic minimax risk (1.2) is consistent with the regular case (van der Vaart and Wellner, 1996); moreover, it is also consistent with previous work by Song (2014a) who studies a more restrictive class of irregular parameters.

We also present a general procedure of constructing optimal plug-in estimators. An optimal plug-in estimator is of the form $\phi(\hat{\theta}_n + \hat{u}_n/r_n)$, where $\hat{\theta}_n$ is an efficient estimator of θ_0 usually available from efficient estimation literature, and $\hat{u}_n$ is a correction term that depends on the particular loss function ℓ. It is interesting to note that optimality is preserved under simple plug-in for differentiable maps (van der Vaart, 1991b), but in general not for nondifferentiable ones due to the presence of the correction term $\hat{u}_n$ – i.e. $\hat{u}_n$ equals zero when ϕ is differentiable but may be nonzero otherwise. Heuristically, the need of the correction term $\hat{u}_n$ arises from the fact that the simple plug-in estimator $\phi(\hat{\theta}_n)$ may have undesirably high risk at θ_0 where ϕ is nondifferentiable. By adding a correction term, one is able to reduce the risk under the worst case scenario. As an illustration, we apply the construction procedure to the estimation of the effect of Vietnam veteran status on the quantiles of civilian earnings. In the application, the structural quantile functions of earnings exhibit local nonmonotonicity, especially for veterans. Nonetheless, by estimating the closest monotonically increasing functions to the population quantile processes, we are able to resolve this issue and provide locally asymptotically minimax plug-in estimators.

There has been extensive study on optimal estimation of regular parameters (Ibragimov and Has'minskii, 1981; Bickel et al., 1993; Lehmann and Casella, 1998). The best

[1] For example, if P is parametrized as $\theta \mapsto P_\theta$ where θ belongs to an open set $\Theta \subset \mathbf{R}^k$, one typically considers local parametrization $h \mapsto P_{\theta_0 + h/\sqrt{n}}$ with local parameter h ranging over the whole space $\mathbf{R}^k$. We shall have a formal definition of H in Section 2.2.

known optimality results are characterized by the convolution theorems and the local asymptotic minimax theorems (Hájek, 1970; Hájek, 1972; Le Cam, 1972; Koshevnik and Levit, 1976; Levit, 1978; Pfanzagl and Wefelmeyer, 1982; Begun et al., 1983; Millar, 1983, 1985; Chamberlain, 1987; van der Vaart, 1988b; van der Vaart and Wellner, 1990; van der Vaart, 1991a). However, little work has been done on nondifferentiable parameters. Blumenthal and Cohen (1968a,b) considered minimax estimation of the maximum of two translation parameters and pointed out the link between biased estimation and nondifferentiability of the parameter. Hirano and Porter (2012) formally established the connection between differentiability of parameters and possibility of regular, mean unbiased and quantile unbiased estimation, and emphasized the need for alternative optimality criteria when the parameters of interest are nondifferentiable. Chernozhukov et al. (2013) considered estimation of intersection bounds in terms of median-bias-corrected criterion. The work by Song (2014a,b) is mostly closely related to ours. By restricting the parameter of interest to be a composition of a real valued Lipschitz function having a finite set of nondifferentiability points and a translation-scale equivariant real-valued map, Song (2014a,b) was able to establish local asymptotic minimax estimation within the class of arbitrary estimators. In present paper, we consider a much wider class of parameters at the expense of restricting the competing estimators to be of a plug-in form. We note also that for differentiable ϕ, the optimality of the plug-in principle has been established by van der Vaart (1991b).

The remainder of the paper is structured as follows. Section 2 formally introduces the setup, presents a convolution theorem for efficient estimation of θ that will be essential for later discussion, and specifies the suitable version of local asymptotic minimaxity criterion for our purposes. In Section 3 we derive the minimax lower bound for the class of plug-in estimators, and then present a general construction procedure. Section 4 applies the theory to the estimation of the effect of Vietnam veteran status on the quantiles of civilian earnings. Section 5 concludes. All proofs are collected in Appendices.

2 Setup, Convolution and Minimaxity

In this section, we formally set up the problem under consideration, present a convolution theorem for the estimation of θ, and establish the optimality criterion that will be employed to assess the statistical performance of plug-in estimators $\phi(\hat{\theta}_n)$.

2.1 Setup and Notation

In order to accommodate applications such as incomplete auction models and estimation under shape restrictions, we must allow for both the parameter θ_0 and the map ϕ to take values in possibly infinite dimensional spaces; see Examples 2.3 and 2.4 below. We therefore impose the general requirement that $\theta_0 \in \mathbb{D}_\phi$ and $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$ for $\mathbb{D}$ and $\mathbb{E}$ Banach spaces with norms $\|\cdot\|_\mathbb{D}$ and $\|\cdot\|_\mathbb{E}$ respectively, and $\mathbb{D}_\phi$ the domain of ϕ.

The estimator $\hat{\theta}_n$ is assumed to be an arbitrary map of the sample $\{X_i\}_{i=1}^n$ into the domain of ϕ. Thus, the distributional convergence in our context is understood to be in the Hoffman-Jørgensen sense and expectations throughout should be interpreted as outer expectations (van der Vaart and Wellner, 1996), though we obviate the distinction in the notation.

We introduce notation that is recurrent in this paper. For a set T, we denote the space of bounded functions on T by

$$\ell^\infty(T) \equiv \{f : T \to \mathbf{R} : \|f\|_\infty < \infty\} \ , \quad \|f\|_\infty \equiv \sup_{t \in T} |f(t)| \ , \tag{2.1}$$

which is a Banach space under the norm $\|\cdot\|_\infty$. If $(T, \mathcal{M}, \mu)$ is a measure space, we define for $1 \leq p < \infty$,

$$L^p(T, \mathcal{M}, \mu) \equiv \{f : T \to \mathbf{R} : \|f\|_{L^p} < \infty\} \ , \quad \|f\|_{L^p} \equiv \{\int |f|^p \, d\mu\}^{1/p} \ , \tag{2.2}$$

which is a Banach space under the norm $\|\cdot\|_{L^p}$. If the underlying σ-algebra $\mathcal{M}$ or the measure μ is understood without confusion, we also write $L^p(T, \mu)$ or $L^p(T)$. If (T, d) is a metric space, we define

$$\mathrm{BL}_1(T) \equiv \{f : T \to \mathbf{R} : \sup_{t \in T} |f(t)| \leq 1 \text{ and } |f(t_1) - f(t_2)| \leq d(t_1, t_2)\} \ , \tag{2.3}$$

that is, $\mathrm{BL}_1(T)$ is the set of all Lipschitz functions whose level and Lipschitz constant is bounded by 1. For two sets A and B, we write $A \subset_f B$ to signify that A is a finite subset of B. For a finite set $\{g_1, \ldots, g_m\}$, we write $g^m \equiv (g_1, \ldots, g_m)^\intercal$. Lastly, we define $K_\lambda^m \equiv \{x \in \mathbf{R}^m : \|x\| \leq \lambda\}$ for $\lambda > 0$.

2.1.1 Examples

To illustrate the applications of our framework, we begin by presenting some examples that arise in the econometrics and statistics literature. We shall revisit these examples later on as we develop our theory. To highlight the essential ideas and for ease of exposition, we base our discussion on simplifications of well known examples. The general case can be handled analogously.

In the treatment effect literature one might be interested in estimating the maximal treatment effect. Our first example has been considered in Hirano and Porter (2012) and Song (2014a,b).

Example 2.1 (Best Treatment)**.** Let $X = (X^{(1)}, X^{(2)})^\intercal \in \mathbf{R}^2$ be a pair of potential outcomes under two treatments. Consider the problem of estimating the parameter

$$\phi(\theta_0) = \max\{\mathbb{E}[X^{(1)}], \mathbb{E}[X^{(2)}]\} \ . \tag{2.4}$$

One can think of $\phi(\theta_0)$ as the expected outcome under the best treatment. In this case, $\theta_0 = (\mathbb{E}[X^{(1)}], \mathbb{E}[X^{(2)}])^\intercal$, $\mathbb{D} = \mathbf{R}^2$, $\mathbb{E} = \mathbf{R}$, and $\phi : \mathbf{R}^2 \to \mathbf{R}$ is given by $\phi(\theta) = \max(\theta^{(1)}, \theta^{(2)})$. Parameters of this type are essential in characterizing optimal decision rules in dynamic treatment regimes which, as opposed to classical treatment, incorporate heterogeneity across both individuals and time (Murphy, 2003). We note that the functional form of (2.4) is also related to the study of bounds of treatment effects under monotone instruments (Manski and Pepper, 2000, 2009). Minimax estimation of $\phi(\theta)$ when $X^{(1)}$ and $X^{(2)}$ are independent normal random variables with equal variances has been studied in Blumenthal and Cohen (1968a,b). ■

Partial identification is an inherent feature of statistical analysis based on interval censored data. In these settings, one might still want to estimate identified features of the model under consideration. Our second example is based on Manski and Tamer (2002) who study inference on regressions with interval data on a regressor or outcome.

Example 2.2 (Interval Regression Model)**.** Let $Y \in \mathbf{R}$ be a random variable generated by

$$Y = \alpha + \beta W + \epsilon \ ,$$

where $W \in \{-1, 0, 1\}$ is a discrete random variable, and $E[\epsilon|W] = 0$. Suppose that Y is unobservable but there exist (Y_l, Y_u) such that $Y_l \leq Y \leq Y_u$ almost surely. Let $\vartheta = (\alpha, \beta)^\intercal$ and $Z = (1, W)^\intercal$. Then the identified region for ϑ is

$$\Theta_0 \equiv \{\vartheta \in \mathbf{R}^2 : E[Y_l|Z] \leq Z^\intercal \vartheta \leq E[Y_u|Z]\} \ .$$

Interest often centers on either the maximal value of a particular coordinate of ϑ or the maximal value of the conditional expectation $E[Y|W]$ at a specified value of the covariates, both of which can be expressed as

$$\sup_{\vartheta \in \Theta_0} \lambda^\intercal \vartheta \ , \tag{2.5}$$

for some known $\lambda \equiv (\lambda^{(1)}, \lambda^{(2)})^\intercal \in \mathbf{R}^2$. Let $\theta_0 \equiv (P(W = -1), P(W = 1))^\intercal$. It is shown in Appendix B that the analysis of (2.5) reduces to examining terms of the form[2]

$$\phi(\theta_0) = \max\{\psi(\theta_0), 0\} \ , \tag{2.6}$$

where for each $\theta = (\theta^{(1)}, \theta^{(2)})^\intercal \in \mathbf{R}^2$, $\psi(\theta)$ is defined by

$$\psi(\theta) = \lambda^{(1)} \frac{\theta^{(1)} + \theta^{(2)}}{\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2} + \lambda^{(2)} \frac{\theta^{(1)} - \theta^{(2)}}{\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2} \ .$$

In this example, $\mathbb{D} = \mathbf{R}^2$, $\mathbb{E} = \mathbf{R}$ and $\phi : \mathbf{R}^2 \to \mathbf{R}$ satisfies $\phi(\theta) = \max\{\psi(\theta), 0\}$ with $\psi(\theta)$ defined as above. The functional form of ϕ here is common in a class of partially identified models (Beresteanu and Molinari, 2008; Bontemps et al., 2012; Chandrasekhar et al., 2012; Kaido and Santos, 2013; Kaido, 2013; Kline and Santos, 2013). ■

The next example presents a nondifferentiable function which appears as an identification bound on the distribution of valuations in an incomplete model of English auctions (Haile and Tamer, 2003; Hirano and Porter, 2012).

Example 2.3 (Incomplete Auction Model)**.** In an English auction model with symmetric independent private values, a robust approach of interpreting bidding data proposed by Haile and Tamer (2003) is to assume only that bidders neither bid more than their valuations nor let an opponent win at a price they would be willing to beat. Consider two auctions in which bidders' valuations are i.i.d. draws from F. Let B_i and V_i be bidder i's bid and valuation respectively, and let F_1 and F_2 be the distributions of bids in two auctions. The first assumption implies $B_i \leq V_i$ for all i, which in turn imposes a upper bound of F:[3]

$$F(v) \leq \min\{F_1(v), F_2(v)\} \ .$$

Similarly, by exploiting the assumption that bidders do not let an opponent win at a price below their willingness to pay, one may obtain a lower bound on F. For simplicity, we consider only the upper bound which we write as

$$\phi(\theta_0)(v) = \min\{F_1(v), F_2(v)\} \ . \tag{2.7}$$

In this example, $\theta_0 = (F_1, F_2)$, $\mathbb{D} = \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$, $\mathbb{E} = \ell^\infty(\mathbf{R})$ and $\phi : \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R}) \to \ell^\infty(\mathbf{R})$ satisfies $\phi(\theta)(v) \equiv \max\{\theta^{(1)}(v), \theta^{(2)}(v)\}$. ■

[2]Here we work with $\phi(\theta_0)$ for simplicity and ease of exposition.

[3]Haile and Tamer (2003) actually exploit order statistics of bids in order to obtain tighter bounds on F.

Our final example involves a map that monotonizes estimators in linear quantile regressions. Being estimated in pointwise manner, the quantile regression processes need not be monotonically increasing (Bassett and Koenker, 1982; He, 1997). This problem can be fixed by considering the closest monotonically increasing function.[4]

Example 2.4 (Quantile Functions without Crossing)**.** Let $Y \in \mathbf{R}$ and $Z \in \mathbf{R}^d$ be random variables. Consider the linear quantile regression model:

$$\beta(\tau) \equiv \arg\min_{\beta \in \mathbf{R}^d} E[\rho_\tau(Y - Z'\beta)] \ ,$$

where $\rho_\tau(u) \equiv u(\tau - 1\{u \geq 0\})$. Let $\mathcal{T} \equiv [\epsilon, 1-\epsilon]$ for some $\epsilon \in (0, 1/2)$ and $\theta_0 \equiv c'\beta(\cdot) : \mathcal{T} \to \mathbf{R}$ be the quantile regression process, for fixed $Z = c$. Under misspecification, θ_0 need not be monotonically increasing. In order to avoid the quantile crossing problem, we may instead consider projecting θ_0 onto the set of monotonically increasing functions – i.e. the closest monotonically increasing function to θ_0:

$$\phi(\theta_0) = \Pi_\Lambda \theta_0 \equiv \arg\min_{\lambda \in \Lambda} \|\lambda - \theta_0\|_{L^2} \ , \tag{2.8}$$

where Λ be the set of monotonically increasing functions in $L^2(\mathcal{T}, \nu)$ with ν the Lebesgue measure on $\mathcal{T}$, and Π_Λ is the metric projection onto Λ – i.e. the mapping that assigns every point in $L^2(\mathcal{T})$ with the closest point in Λ.[5] In this example, $\mathbb{D} = L^2(\mathcal{T})$, $\mathbb{E} = \Lambda$ and $\phi : L^2(\mathcal{T}) \to \Lambda$ is defined by $\phi(\theta) = \Pi_\Lambda \theta$. We note that the metric projection approach introduced here can in fact handle a larger class of estimation problems under shape restrictions; see Remark 2.1. ■

Remark 2.1. Let $\theta_0 : \mathcal{T} \to \mathbf{R}$ be a unknown real valued function where $\mathcal{T} = [a, b]$ with $-\infty < a < b < \infty$. Then one may monotonize θ_0 by considering the nearest monotonically increasing function $\phi(\theta_0) \equiv \Pi_\Lambda \theta_0$ where $\Lambda \subset L^2(\mathcal{T})$ is the set of increasing functions. More generally, one may take Λ to be a closed and convex set of functions satisfying certain shape restrictions such as convexity and homogeneity. Then the projection $\Pi_\Lambda \theta_0$ of θ_0 onto Λ is the closest function to θ_0 with desired shape restrictions. ■

2.2 The Convolution Theorem

In this section, before delving into the discussion of the defining ingredient ϕ, we formalize basic regularity assumptions and then present a convolution theorem for the estimation of θ_0, which in turn will be employed when deriving the asymptotic minimax lower bound for the estimation of $\phi(\theta_0)$.

Following the literature on limits of experiments (Blackwell, 1951; Le Cam, 1972; van der Vaart, 1991a), we consider a sequence of experiments $\mathcal{E}_n \equiv (\mathscr{X}_n, \mathcal{A}_n, \{P_{n,h} : h \in H\})$, where $(\mathscr{X}_n, \mathcal{A}_n)$ is a measurable space, and $P_{n,h}$ is a probability measure on $(\mathscr{X}_n, \mathcal{A}_n)$, for each $n \in \mathbf{N}$ and $h \in H$ with H a subspace of some Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle_H$ and induced norm $\| \cdot \|_H$. We observe a sample $X_1, \ldots, X_n$ that is jointly distributed according to some $P_{n,h}$. This general framework

[4] Alternatively, Chernozhukov et al. (2010) propose employing a sorting operator to monotonize possibly nonmonotone estimators.

[5] The set Λ is closed and convex so that the metric projection Π_Λ exists and is unique; see Appendix B for detailed discussion.

allows us to consider non i.i.d. models (Ibragimov and Has'minskii, 1981; van der Vaart, 1988b; van der Vaart and Wellner, 1990) as well as common i.i.d. setup. We confine our attention to the family of probability measures $\{P_{n,h} : h \in H\}$ possessing local asymptotic normality; see Assumption 2.1(ii).[6] This is perhaps the most convenient class to begin with in the literature of efficient estimation, since mutual contiguity implied by local asymptotic normality allows us, by Le Cam's third lemma, to deduce weak limits along sequence $\{P_{n,h}\}_{n=1}^{\infty}$ from that under the fixed sequence $\{P_{n,0}\}_{n=1}^{\infty}$ – usually thought of as the underlying truth. Formally, we impose

Assumption 2.1 *(i) The set H is a subspace of some separable Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$ and induced norm $\| \cdot \|_H$.*

(ii) The sequence of experiments $(\mathcal{X}_n, \mathcal{A}_n, \{P_{n,h} : h \in H\})$ is asymptotically normal, i.e.

$$\log \frac{dP_{n,h}}{dP_{n,0}} = \Delta_{n,h} - \frac{1}{2}\|h\|_H^2 \ , \tag{2.9}$$

where $\{\Delta_{n,h} : h \in H\}$ is a stochastic process which converges to $\{\Delta_h : h \in H\}$ marginally under $\{P_{n,0}\}$,[7] with $\{\Delta_h : h \in H\}$ a Gaussian process having mean zero and covariance function given by $E[\Delta_{h_1}\Delta_{h_2}] = \langle h_1, h_2 \rangle_H$.[8]

Separability as in Assumption 2.1(i) is only a minimal requirement in practice, while linearity is standard although not entirely necessary.[9] The essence of Assumption 2.1(ii) is that the sequence of experiments $\mathcal{E}_n$ can be asymptotically represented by a Gaussian shift experiment. Thus, one may "pass to the limit first", "argue the case for the limiting problem" which has simpler statistical structure, and then translate the results back to the original experiments $\mathcal{E}_n$ (Le Cam, 1972).[10] In the i.i.d. case, Assumption 2.1(ii) is guaranteed by the so-called differentiability in quadratic mean; see Remark 2.2.

Regularity conditions on the parameter θ and an estimator $\hat{\theta}_n$ are imposed as follows. In our setup, we recognize θ as a map $\theta : \{P_{n,h}\} \to \mathbb{D}$ and write $\theta_n(h) \equiv \theta(P_{n,h})$.

Assumption 2.2 *The parameter $\theta : \{P_{n,h}\} \to \mathbb{D}_\phi \subset \mathbb{D}$, where $\mathbb{D}$ is a Banach space with norm $\| \cdot \|_{\mathbb{D}}$, is regular, i.e. there exists a continuous linear map $\theta_0' : H \to \mathbb{D}$ such that for every $h \in H$,*

$$r_n\{\theta_n(h) - \theta_n(0)\} \to \theta_0'(h) \ \ as \ \ n \to \infty \ , \tag{2.10}$$

for a sequence of $\{r_n\}$ with $r_n \to \infty$ as $n \to \infty$.

[6] Our results in fact extend to models having local asymptotic mixed normality; see Jeganathan (1981, 1982) and van der Vaart (1998, Section 9.6).

[7] That is, for any finite set $I \subset H$, $(\Delta_{n,h} : h \in I) \xrightarrow{L} (\Delta_h : h \in I)$ under $\{P_{n,0}\}$.

[8] Here, $dP_{n,0}$ and $dP_{n,h}$ can be understood as densities of $P_{n,0}$ and $P_{n,h}$ with respect to some σ-finite measure μ_n, respectively. Fortunately, the log ratio above is independent of the choice of μ_n; see van der Vaart (1998, p.189-91).

[9] In fact, H can be relaxed to be a convex cone; see van der Vaart and Wellner (1996) and van der Vaart (1989).

[10] From a technical level, for any finite set $I \subset H$, weak convergence of likelihoods in Assumption 2.1(ii) is equivalent to convergence in terms of Le Cam's deficiency distance (Le Cam, 1972, 1986).

Assumption 2.3 $\hat{\theta}_n : \{X_i\} \to \mathbb{D}_\phi$ *is regular, i.e. there is a fixed tight random variable* $\mathbb{G} \in \mathbb{D}$ *such that for any* $h \in H$*,*

$$r_n\{\hat{\theta}_n - \theta_n(h)\} \overset{L_{n,h}}{\to} \mathbb{G} \text{ in } \mathbb{D} \ , \tag{2.11}$$

where $\overset{L_{n,h}}{\to}$ *denotes weak convergence under* $\{P_{n,h}\}$*.*

Assumption 2.2, which dates back to Pfanzagl and Wefelmeyer (1982), is essentially a Hadamard differentiability requirement; see Remark 2.3. Our optimality analysis shall extend from Hadamard differentiable parameters to a class of (Hadamard) directionally differentiable parameters. The derivative $\theta_0' : H \to \mathbb{D}$ is crucial in determining the efficiency bound for estimating θ. If $\mathbb{D} = \mathbf{R}^m$, the derivative $\theta_0' : H \to \mathbf{R}^m$ uniquely determines through the Riesz representation theorem a $m \times 1$ vector $\tilde{\theta}_0$ of elements in the completion $\overline{H}$ of H such that $\theta_0'(h) = \langle \tilde{\theta}_0, h \rangle$ for all $h \in H$. The matrix $\Sigma_0 \equiv \langle \tilde{\theta}_0, \tilde{\theta}_0^{\intercal} \rangle$ is called the efficiency bound for θ. For general $\mathbb{D}$, the efficiency bound is characterized through the topological dual space $\mathbb{D}^*$ of $\mathbb{D}$ (Bickel et al., 1993); see Theorem 2.1.

Assumption 2.3 means that $\{\hat{\theta}_n\}$ is asymptotically equivariant in law for estimating $\theta_n(h)$, or put it another way, the limiting distribution of $\{\hat{\theta}_n\}$ is robust to "local perturbations" $\{P_{n,h}\}$ of the "truth" $\{P_{n,0}\}$. In this way it restricts the class of plug-in estimators we consider. For instance, superefficient estimators such as Hodges' estimator and shrinkage estimators are excluded from our setup (Le Cam, 1953; Huber, 1966; Hájek, 1972; van der Vaart, 1992). Finally, we note that while regularity of θ, as ensured by Assumption 2.2, is necessary for Assumption 2.3 to hold (Hirano and Porter, 2012), it is in general not sufficient unless the model is parametric (Bickel et al., 1993).

Assumptions 2.1, 2.2 and 2.3 together place strong restrictions on the structure of the asymptotic distribution of $\hat{\theta}_n$. In particular, for every $\hat{\theta}_n$ satisfying the above regularity conditions, its weak limit can be represented as the efficient Gaussian random variable plus an independent noise term, as illustrated in the following convolution theorem taken directly from van der Vaart and Wellner (1990). The derivative $\theta_0' : H \to \mathbb{D}$ as a continuous linear map has an adjoint map $\theta_0'^* : \mathbb{D}^* \to \overline{H}$ satisfying $d^*\theta_0'(h) = \langle \theta_0'^* d^*, h \rangle_H$ for all $d^* \in \mathbb{D}^*$; that is, $\theta_0'^*$ maps the dual space $\mathbb{D}^*$ of $\mathbb{D}$ into $\overline{H}$.

Theorem 2.1 (Hájek-Le Cam Convolution Theorem) *Let* $(\mathscr{X}_n, \mathcal{A}_n, \{P_{n,h} : h \in H\})$ *be a sequence of statistical experiments, and* $\hat{\theta}_n$ *be an estimator for the parameter* $\theta : \{P_{n,h}\} \to \mathbb{D}$*. Suppose that Assumptions 2.1, 2.2 and 2.3 hold. It follows that*[11]

$$\mathbb{G} \overset{d}{=} \mathbb{G}_0 + \mathbb{U} \ , \tag{2.12}$$

where $\mathbb{G}_0$ *is a tight Gaussian random variable in* $\mathbb{D}$ *satisfying* $d^*\mathbb{G}_0 \sim \mathcal{N}(0, \|\theta_0'^* d^*\|_H^2)$ *for every* $d^* \in \mathbb{D}^*$*, and* $\mathbb{U}$ *is a tight random variable in* $\mathbb{D}$ *that is independent of* $\mathbb{G}_0$*. Moreover, the support of* $\mathbb{G}_0$ *is* $\overline{\theta_0'(H)}$ *(the closure of* $\{\theta_0'(h) : h \in H\}$ *relative to* $\|\cdot\|_{\mathbb{D}}$*).*[12]

One important implication of Theorem 2.1 is that a regular estimator sequence $\{\hat{\theta}_n\}$ is considered efficient if its limiting law is such that $\mathbb{U}$ is degenerate at 0. In addition, normality being "the best limit" is a result of optimality, rather than an *ex*

[11] The symbol $\overset{d}{=}$ denotes equality in distribution.

[12] The support of $\mathbb{G}_0$ refers to the intersection of all closed subsets $\mathbb{D}_0 \subset \mathbb{D}$ with $P(\mathbb{G}_0 \in \mathbb{D}_0) = 1$.

ante restriction. If ϕ is Hadamard differentiable, then we may conclude immediately that $\phi(\hat{\theta}_n)$ is an efficient estimator for $\phi(\theta_0)$ if $\hat{\theta}_n$ is for θ_0 (van der Vaart, 1991b). When ϕ is Hadamard directionally differentiable only, however, we have to base our optimality analysis within the class of irregular estimators because no regular estimators exist in this context (Hirano and Porter, 2012). As a result, the convolution theorem is not available in general, which motivates the optimality analysis in terms of asymptotic minimax criterion.

Remark 2.2. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample with common distribution P that is known to belong to a collection $\mathcal{P}$ of Borel probability measures, and let $\{P_t : t \in (0, \epsilon)\} \subset \mathcal{P}$ with $P_0 = P$ be a submodel such that

$$\int \left[\frac{dP_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2} h dP^{1/2} \right]^2 \to 0 \text{ as } t \downarrow 0 , \tag{2.13}$$

where h is called the score of this submodel. In this situation, we identify $P_{n,h}$ with $\prod_{i=1}^n P_{1/\sqrt{n},h}$ where $\{P_{1/\sqrt{n},h}\}$ is differentiable in quadratic mean with score h, and the set $\dot{\mathcal{P}}^0$ of all score functions thus obtained, which are necessarily elements of $L^2(P)$, will be the index set H, also known as the tangent set of $\mathcal{P}$. It can be shown that the sequence $\{P_{n,h}\}$ satisfies Assumption 2.1(ii) (van der Vaart and Wellner, 1996). ■

Remark 2.3. Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample generated according to some $P \in \mathcal{P}$ where $\mathcal{P}$ is dominated by a σ-finite measure μ. Since $\mathcal{P}$ can be embedded into $L^2(\mu)$ via the mapping $Q \mapsto \sqrt{dQ/d\mu}$, we can obtain a tangent set $\dot{\mathcal{S}}^0$ consisting of Fréchet derivatives of differentiable paths $\{dP_t^{1/2}\}$ in $L^2(\mu)$ (Bickel et al., 1993). Define the continuous linear operator $\dot{\theta}_0 : \dot{\mathcal{S}}^0 \to \mathbb{D}$ by $\dot{\theta}_0(g) \equiv \theta_0'(2g/dP^{1/2})$, then (2.10) can be read as

$$\lim_{t \downarrow 0} t^{-1} \{\theta(dP_t^{1/2}) - \theta(dP^{1/2})\} = \dot{\theta}_0(g) , \tag{2.14}$$

where $\{dP_t^{1/2}\}$ is a curve passing $dP^{1/2}$ with Fréchet derivative $g \equiv \frac{1}{2} h dP^{1/2}$. This is exactly Hadamard differentiability if we view θ as a map from $\{\sqrt{dQ/d\mu} : Q \in \mathcal{P}\} \subset L^2(\mu)$ to the space $\mathbb{D}$. ■

2.3 Local Asymptotic Minimaxity

There are different versions of local asymptotic minimax risk. In this section we briefly review some of these and specify the one that is appropriate for our purposes. For simplicity of exposition, let us confine our attention to the i.i.d. case. Let $\mathcal{P}$ be a collection of probability measures, θ the parameter of interest and ℓ a loss function. In an asymptotic framework, a global minimax principle would imply that an asymptotically best estimator sequence $\{T_n\}$ of θ should be the one for which the quantity

$$\liminf_{n \to \infty} \sup_{P \in \mathcal{P}} E_P[\ell(r_n\{T_n - \theta(P)\})] \tag{2.15}$$

is minimized, where E_P denotes expectation under P, and $r_n \uparrow \infty$ is the rate of convergence for estimating θ. While this version is suitable when $\mathcal{P}$ is parametric, it is somewhat too restrictive for semiparametric or nonparametric models. In addition,

this approach is excessively cautious since we are able to learn about P with arbitrary accuracy as sample size $n \to \infty$ and hence it would be unreasonable to require nice properties of the estimator sequence around regions too far away from the truth (Hájek, 1972; Ibragimov and Has'minskii, 1981; van der Vaart, 1992). The strategy is then to minimize the asymptotic maximum risk over (shrinking) neighborhoods of the truth.

The earliest consideration of *local* asymptotic minimaxity in the literature is perhaps Chernoff (1956), according to whom the idea actually originated from Charles Stein and Herman Rubin. Different variants have been developed since then (Hájek, 1972; Koshevnik and Levit, 1976; Levit, 1978; Chamberlain, 1987), among which an important version is of the form

$$\lim_{a\to\infty} \liminf_{n\to\infty} \sup_{P\in V_{n,a}} E_P[\ell(r_n\{T_n - \theta(P)\})] \ , \tag{2.16}$$

where $V_{n,a}$ shrinks to the truth as $n \to \infty$ for each fixed $a \in \mathbf{R}$ and spans the whole parameter space as $a \to \infty$ for each fixed $n \in \mathbf{N}$ (Ibragimov and Has'minskii, 1981; Millar, 1983). For instance, Begun et al. (1983) and van der Vaart (1988b) take $V_{n,a}$ to be:

$$V_{n,a} = \{Q \in \mathcal{P} : r_n d_H(Q,P) \le a\} \ , \quad d_H(Q,P) \equiv \left[\int (dQ^{1/2} - dP^{1/2})^2\right]^{1/2} \ .$$

However, the above neighborhood versions may invite two problems. First, the neighborhoods might be too large so that the sharp lower bounds are infinite. This is more easily seen in the Hellinger ball version. As pointed out by van der Vaart (1988b, p.32), one may pick $Q_n \in V_n(P,a)$ for each $n \in \mathbf{N}$ such that $\prod_{i=1}^n Q_n$ is not contiguous to $\prod_{i=1}^n P$ (Oosterhoff and Zwet, 1979, Theorem 1), which in turn implies that $r_n\{T_n - \theta\}$ escapes to "infinity" under $\prod_{i=1}^n Q_n$ (Lehmann and Romano, 2005, Theorem 12.3.2). Second, when it comes to the construction of an optimal estimator, one typically has to establish uniform convergence over the neighborhoods, which may be impossible if the neighborhoods are "too big".

In this paper, we shall consider local asymptotic minimax risk over smaller neighborhoods – more precisely, neighborhoods that consist of finite number of distributions – as in van der Vaart (1988b, 1989, 1998) and van der Vaart and Wellner (1990, 1996):

$$\sup_{I\subset_f \dot{\mathcal{P}}^0} \liminf_{n\to\infty} \sup_{h\in I} E_{P_{n,h}}[\ell(r_n\{T_n - \theta(P_{n,h})\})] \ , \tag{2.17}$$

where the first supremum is taken over all finite subsets I in the tangent set $\dot{\mathcal{P}}^0$ as defined in Remark 2.2, and $\{P_{n,h}\}$ is a differentiable path with score h. This resolves the aforementioned concerns as well as two subtleties that are worth noting here. First, it is necessary to take supremum over neighborhoods of the truth (the second supremum) in order to obtain robust finite sample approximation and as a result rule out superefficient estimators, while the first supremum is needed to remove the uncertainty of the neighborhoods.[13] Second, the local nature of the risk may be translated to the global one if one replaces the second supremum with $\sup_{h\in\dot{\mathcal{P}}^0}$ and ignore the first supremum, so that we are back to the aforementioned uniformity issue. Another possibility is to consider finite dimensional submodels; see Remark 2.4.

[13]The role played by $\sup_{I\subset_f \dot{\mathcal{P}}^0}$ is the same as that by $\lim_{a\to\infty}$ in display (2.16).

Remark 2.4. As another approach to circumvent the contiguity and uniformity concerns aforementioned, van der Vaart (1988b) considers a version of asymptotic minimaxity based on finite dimensional submodels. Let $h_1, \ldots, h_m \in \dot{\mathcal{P}}^0$ be linearly independent and $\{P^m_{n,\lambda}\}_{n=1}^{\infty}$ a differentiable path with score $\sum_{j=1}^m \lambda_j h_j$ for each fixed $\lambda \in \mathbf{R}^m$. As λ ranges over $\mathbf{R}^m$, we obtain a full description of local perturbations of some parametric submodel. Then one may consider the following:

$$\sup_{H_m} \lim_{a\to\infty} \liminf_{n\to\infty} \sup_{\|\lambda\|\le a} E_{P^m_{n,\lambda}}[\ell(r_n\{T_n - \theta(P^m_{n,\lambda})\})] \ , \tag{2.18}$$

where the first supremum is taken over all finite dimensional subspaces $H_m \subset \dot{\mathcal{P}}^0$ spanned by $h_1, \ldots, h_m$. The same approach has been employed by van der Vaart (1988b, 1989) to obtain generalized convolution theorems for weakly regular estimators. We note however that this version of local asymptotic minimaxity is equivalent to (2.17) in the sense that they yield the same lower bound that is attainable and hence induce the same optimal plug-in estimators. This is essentially because for any parametric submodel $\mathcal{P}^m$ with scores $h_1, \ldots, h_m$, the expansion of the log likelihood ratio (2.9) holds uniformly over $\lambda \in K$ with K any compact set in $\mathbf{R}^m$ (Bickel et al., 1993, Proposition 2.1.2). ■

3 Optimal Plug-in Estimators

Building on the ingredients established for θ in previous section, we now proceed to investigate optimal plug-in estimators of $\phi(\theta)$. To begin with, we first review the notion of Hadamard directional differentiability, then establish the minimax lower bound for the class of plug-in estimators, and finally show the attainability by presenting a general procedure of constructing optimal plug-in estimators.

3.1 Hadamard Directional Differentiability

A common feature of the examples introduced in Section 2.1.1 is that there exist points $\theta \in \mathbb{D}$ at which the map $\phi : \mathbb{D} \to \mathbb{E}$ is not differentiable. Nonetheless, at all such θ at which differentiability is lost, ϕ actually remains directionally differentiable. This is most easily seen in Examples 2.1 and 2.2, in which the domain of ϕ is a finite dimensional space. In order to address Examples 2.3 and 2.4, however, a notion of directional differentiability that is suitable for more abstract spaces $\mathbb{D}$ is necessary. Towards this end, we follow Shapiro (1990) and define

Definition 3.1. Let $\mathbb{D}$ and $\mathbb{E}$ be Banach spaces equipped with norms $\|\cdot\|_{\mathbb{D}}$ and $\|\cdot\|_{\mathbb{E}}$ respectively, and $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$. The map ϕ is said to be *Hadamard directionally differentiable* at $\theta \in \mathbb{D}_\phi$ if there is a map $\phi'_\theta : \mathbb{D} \to \mathbb{E}$ such that:

$$\lim_{n\to\infty} \left\| \frac{\phi(\theta + t_n z_n) - \phi(\theta)}{t_n} - \phi'_\theta(z) \right\|_{\mathbb{E}} = 0 \ , \tag{3.1}$$

for all sequences $\{z_n\} \subset \mathbb{D}$ and $\{t_n\} \subset \mathbf{R}_+$ such that $t_n \downarrow 0$, $z_n \to z$ as $n \to \infty$ and $\theta + t_n z_n \in \mathbb{D}_\phi$ for all n.

As various notions of differentiability in the literature, Hadamard directional differentiability can be understood by looking at the restrictions imposed on the approximating

map (i.e. the derivative) and the way the approximation error is controlled (Averbukh and Smolyanov, 1967, 1968). Specifically, let

$$\text{Rem}_\theta(z) \equiv \phi(\theta + z) - \{\phi(\theta) + \phi'_\theta(z)\} \ , \tag{3.2}$$

where $\phi(\theta)+\phi'_\theta(z)$ can be viewed as the first order approximation of $\phi(\theta+z)$. Hadamard directional differentiability of ϕ then amounts to requiring the approximation error $\text{Rem}_\theta(z)$ satisfy that $\text{Rem}_\theta(tz)/t$ tends to zero uniformly in $z \in K$ for any compact set K – i.e.

$$\sup_{z\in K} \| \frac{\text{Rem}_\theta(tz)}{t} \|_{\mathbb{E}} \to 0 \ , \text{ as } t \downarrow 0 \ .$$

However, unlike Hadamard differentiability that requires the approximating map ϕ'_θ be linear and continuous, linearity of the directional counterpart is often lost though the continuity is automatic (Shapiro, 1990). In fact, linearity of the derivative is the exact gap between these two notions of differentiability.

The way that Hadamard directional differentiability controls the approximation error ensures the validity of the Delta method, which we exploit in our asymptotic analysis. Moreover, the chain rule remains valid for compositions of Hadamard directional differentiable maps; see Remark 3.1.[14] We note also that though Definition 3.1 is adequate for our purposes in this paper, there is a *tangential* version of Hadamard directional differentiability, which restricts the domain of the derivative ϕ'_{θ_0} to be a subset of $\mathbb{D}$.

Remark 3.1. Suppose that $\psi : \mathbb{B} \to \mathbb{D}_\phi \subset \mathbb{D}$ and $\phi : \mathbb{D}_\phi \to \mathbb{E}$ are Hadamard directionally differentiable at $\vartheta \in \mathbb{B}$ and $\theta \equiv \psi(\vartheta) \in \mathbb{D}_\phi$ respectively, then $\phi \circ \psi : \mathbb{B} \to \mathbb{E}$ is Hadamard directionally differentiable (Shapiro, 1990) at ϑ with derivative $\phi'_\theta \circ \psi'_\vartheta : \mathbb{B} \to \mathbb{E}$. Thus, if $\theta : \{P_{n,h}\} \to \mathbb{D}_\phi$ is not regular but $\theta(P_{n,h}) = \psi(\vartheta(P_{n,h}))$ for some parameter $\vartheta : \{P_{n,h}\} \to \mathbb{B}$ admitting a regular estimator $\hat{\vartheta}_n$ and a Hadamard directionally differentiable map ψ, then the results in this paper may be applied with $\tilde{\phi} \equiv \phi \circ \psi$, $\tilde{\theta}(P_{n,h}) \equiv \vartheta(P_{n,h})$, and $\hat{\vartheta}_n$ in place of ϕ, $\theta(P_{n,h})$ and $\hat{\theta}_n$ respectively. ■

3.1.1 Examples Revisited

We next verify Hadamard directional differentiability of the maps in the examples introduced in Section 2.1.1, and hence show that they indeed fall into our setup. The first example is straightforward.

Example 2.1 (Continued). Let $j^* = \arg\max_{j\in\{1,2\}} \theta^{(j)}$. For any $z = (z^{(1)}, z^{(2)})' \in \mathbf{R}^2$, simple calculations reveal that $\phi'_\theta : \mathbf{R}^2 \to \mathbf{R}$ is given by

$$\phi'_\theta(z) = \begin{cases} z^{(j^*)} & \text{if } \theta^{(1)} \neq \theta^{(2)} \\ \max\{z^{(1)}, z^{(2)}\} & \text{if } \theta^{(1)} = \theta^{(2)} \end{cases} \ . \tag{3.3}$$

Note that ϕ'_θ is nonlinear precisely when Hadamard differentiability is not satisfied. ■

[14] In fact, by slight modifications of the arguments employed in Averbukh and Smolyanov (1968), one can show that Hadamard directional differentiability is the weakest directional differentiability that satisfies the chain rule, just as Hadamard differentiability is the weakest differentiability that does the same job.

Example 2.2 (Continued). In this example, by the chain rule (see Remark 3.1) it is easy to verify that

$$\phi'_\theta(z) = \psi'_\theta(z)1\{\psi(\theta) > 0\} + \max\{\psi'_\theta(z), 0\}1\{\psi(\theta) = 0\} \ , \tag{3.4}$$

where $1\{\cdot\}$ denotes the indicator function, and

$$\begin{aligned}\psi'_\theta(z) =& \frac{[\lambda^{(1)}(z^{(1)} + z^{(2)}) + \lambda^{(2)}(z^{(1)} - z^{(2)})][\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2]}{\left[\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2\right]^2} \\ &- \frac{[\lambda^{(1)}(\theta^{(1)} + \theta^{(2)}) + \lambda^{(2)}(\theta^{(1)} - \theta^{(2)})][z^{(1)} + z^{(2)} - 2(\theta^{(2)} - \theta^{(1)})(z^{(2)} - z^{(1)})]}{\left[\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2\right]^2} \ .\end{aligned}$$

Clearly, the directional derivative ϕ'_θ is nonlinear at θ with $\psi(\theta) = 0$. ■

Example 2.3 and 2.4 are more involved in that the domain and range of ϕ are both infinite dimensional.

Example 2.3 (Continued). Let $B_1 = 1\{x : \theta^{(1)}(x) > \theta^{(2)}(x)\}$, $B_2 = 1\{x : \theta^{(2)}(x) > \theta^{(1)}(x)\}$ and $B_0 = 1\{x : \theta^{(1)}(x) = \theta^{(2)}(x)\}$. Then it is not hard to show that ϕ is Hadamard directionally differentiable at any $\theta \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ satisfying for any $z \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$,

$$\phi'_\theta(z) = z^{(1)}1_{B_1} + z^{(2)}1_{B_2} + \max\{z^{(1)}, z^{(2)}\}1_{B_0} \ . \tag{3.5}$$

Here, nonlinearity occurs when the set of points at which $\theta^{(1)}$ and $\theta^{(2)}$ are equal is not empty, implying Hadamard directional differentiability. ■

Example 2.4 (Continued). For a set $A \subset L^2(\mathcal{T})$, denote the closed linear span of A by $[A]$, and define the complement $A^\perp$ of A by $A^\perp \equiv \{z \in L^2(\mathcal{T}) : \langle z, \lambda\rangle = 0 \text{ for all } \lambda \in A\}$. Lemma B.4 shows that Π_Λ is Hadamard directionally differentiable at every $\theta \in L^2(\mathcal{T})$ and the resulting derivative satisfies for all $z \in L^2(\mathcal{T})$

$$\phi'_\theta(z) = \Pi_{C_\theta}(z) \ , \tag{3.6}$$

where

$$C_\theta = T_{\bar\theta} \cap [\theta - \bar\theta]^\perp \ , \ T_{\bar\theta} = \overline{\bigcup_{\alpha \geq 0} \alpha\{\Lambda - \bar\theta\}} \ , \tag{3.7}$$

with $\bar\theta = \Pi_\Lambda\theta$. Note that C_θ is a closed convex cone, which can be thought of as a local approximation to Λ at θ along the direction perpendicular to the projection residual $\theta - \Pi_\Lambda\theta$. Unlike Fang and Santos (2014), the consideration of nonboundary points $\theta \notin \Lambda$ here is necessitated by the possible misspecification of conditional quantile functions. ■

3.2 The Lower Bounds

As the first step towards establishing the minimax lower bound, we would like to leverage the Delta method for Hadamard directionally differentiable maps (Shapiro, 1991;

Dümbgen, 1993) to derive the weak limits of $r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(h))\}$ under $\{P_{n,h}\}$. This is not a problem in i.i.d. settings since we may write

$$r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(h))\} = r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(0))\} - r_n\{\phi(\theta_n(h)) - \phi(\theta_n(0))\} \ ,$$

and then Delta method can be employed right away in view of the fact that $\theta_n(0)$ is typically a constant. In general, however, we would hope the directional differentiability of ϕ is strong enough to possess uniformity to certain extent.

There are two ways to obtain uniform differentiability in general. One natural way, of course, is to incorporate uniformity into the definition of differentiability (van der Vaart and Wellner, 1996, Theorem 3.9.5). For differentiable maps, continuous differentiability suffices for uniform differentiability; for directionally differentiable ones, unfortunately, continuous differentiability is a rare phenomenon. In fact, one can show by way of example that it is unwise to include uniformity in the definition of Hadamard directional differentiability. The other general principle of obtaining uniformity is to require $\theta_n(0)$ converge sufficiently fast. Following Dümbgen (1993), we take this latter approach and require $\theta_n(0)$ converge in the following manner:

Assumption 3.1 *There are fixed $\theta_0 \in \mathbb{D}_\phi$ and $\Delta \in \theta_0'(H)$ such that as $n \to \infty$,*

$$r_n\{\theta_n(0) - \theta_0\} \to \Delta \ . \tag{3.8}$$

Assumption 3.2 *The map $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$, where $\mathbb{E}$ is a Banach space with norm $\|\cdot\|_\mathbb{E}$, is Hadamard directionally differentiable at θ_0.*

In the i.i.d. setup, Assumption 3.1 is automatically satisfied with $\theta_n(0) = \theta_0 \equiv \theta(P)$, $\Delta = 0$, and $\{r_n\}$ any sequence. Assumption 3.2 simply formalizes the appropriate notion of directional differentiability of ϕ. It is worth noting that directional differentiability is only assumed at θ_0. This Hadamard directional differentiability condition, together with Assumptions 2.2, 2.3, and 3.1, allows us to deduce weak limits of $r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(h))\}$ under $\{P_{n,h}\}$.

Next, minimaxity analysis necessitates the specification of a loss function or a family of loss functions. As recommended by Strasser (1982), we shall consider a collection of loss functions and they are specified as follows:

Assumption 3.3 *The loss function $\ell : \mathbb{E} \to \mathbf{R}^+$ is such that $\ell_M \equiv \ell \wedge M$ is Lipschitz continuous, i.e. for each $M > 0$, there is some constant $C_{\ell,M} > 0$ such that:*

$$|\ell_M(x) - \ell_M(y)| \le C_{\ell,M}\|x - y\|_\mathbb{E} \text{ for all } x, y \in \mathbb{E} \ . \tag{3.9}$$

Assumption 3.3 includes common loss functions such as quadratic loss, absolute loss, and quantile loss but excludes the zero-one loss. We emphasize that the symmetry of ℓ is not needed here. From a technical level, this is because we no longer need Anderson's lemma to derive the lower bound of minimax risk. Moreover, we note that Assumption 3.3 clearly implies continuity of ℓ and Lipschitz continuity if ℓ is bounded.

Given the ability to derive weak limits of $r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(h))\}$, asymptotic normality of $\{P_{n,h}\}$, and a loss function ℓ, we are able to obtain the lower bound of local asymptotic minimax risk as the first main result of this paper.

Theorem 3.1 *Let $(\mathscr{X}_n, \mathcal{A}_n, \{P_{n,h} : h \in H\})$ be a sequence of statistical experiments, and $\hat{\theta}_n$ a map from the data $\{X_i\}_{i=1}^n$ into a set $\mathbb{D}_\phi$. Suppose that Assumptions 2.1, 2.2, 2.3, 3.1, 3.2 and 3.3 hold. Then it follows that*

$$\sup_{I \subset_f H} \liminf_{n\to\infty} \sup_{h\in I} E_{n,h}[\ell(r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(h))\})] \geq \inf_{u\in\mathbb{D}} \sup_{h\in H} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h))\right)\right] , \quad (3.10)$$

where $E_{n,h}$ denotes the expectation evaluated under $P_{n,h}$.

The lower bound takes a minimax form which in fact is consistent with regular cases – i.e. when ϕ is Hadamard differentiable or equivalently ϕ'_{θ_0} is linear, in which the lower bound is given by $E[\ell(\phi'_{\theta_0}(\mathbb{G}_0))]$ provided that ℓ is subconvex (van der Vaart and Wellner, 1996). To see this, note that if ϕ'_{θ_0} is linear, then

$$\inf_{u\in\mathbb{D}} \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] = \inf_{u\in\mathbb{D}} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0) + \phi'_{\theta_0}(u)\right)\right] = E[\ell(\phi'_{\theta_0}(\mathbb{G}_0))] ,$$

where the last step is by Anderson's lemma since ℓ is subconvex and $\phi'_{\theta_0}(\mathbb{G}_0)$ is Gaussian in view of ϕ'_{θ_0} being continuous and linear. Thus, the minimax form in (3.10) is caused entirely by the nonlinearity of ϕ'_{θ_0}. We note also that the lower bound in Theorem 3.1 is consistent with that in Song (2014a) for the special class of parameters studied there.

If the lower bound in (3.10) is infinite, then any estimator is "optimal". One should then change the loss function or work with an alternative optimality criteria so that the problem becomes nontrivial. Given a particular loss function, finiteness of the lower bound hinges on the nature of both the model and the parameter being estimated. For the sake of finiteness of the lower bound, we thus require the derivative ϕ'_{θ_0} satisfy:

Assumption 3.4 *The derivative ϕ'_{θ_0} is Lipschitz continuous, i.e. there exists some constant $C_{\phi'} > 0$ possibly depending on θ_0 such that*

$$\|\phi'_{\theta_0}(z_1) - \phi'_{\theta_0}(z_2)\|_{\mathbb{E}} \leq C_{\phi'}\|z_1 - z_2\|_{\mathbb{D}} \text{ for all } z_1, z_2 \in \mathbb{D}_\phi . \quad (3.11)$$

Assumption 3.4 in fact is satisfied in all of our examples; see Section 3.2.1. The following Lemma shows that Assumption 3.4 ensures finiteness of the lower bound in (3.10) for a class of popular loss functions.

Lemma 3.1 *Let $\ell(\cdot) = \rho(\|\cdot\|_{\mathbb{E}})$ for some nondecreasing lower semicontinuous function $\rho : \mathbf{R}^+ \to \mathbf{R}^+$. If Assumption 3.4 holds and $E[\rho(C_{\phi'}\|\mathbb{G}_0\|_{\mathbb{D}})] < \infty$, then*

$$\inf_{u\in\mathbb{D}} \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] < \infty .$$

The moment condition in Lemma 3.1 is easy to verify in practice when combined with Lipschitz property of ρ (Bogachev, 1998, Theorem 4.5.7) or tail behavior of the CDF of $\|\mathbb{G}_0\|_{\mathbb{E}}$ (Davydov et al., 1998, Proposition 11.6) but by no means necessary. If the lower bound is finite, this would not be a concern in the first place. As another example, if $\mathbb{D}$ is Euclidean, then it suffices that there is some $\delta > 0$ such that

$$\sup_{c\in\mathbf{R}^m} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + c) - \phi'_{\theta_0}(c))^{1+\delta}] < \infty .$$

In cases when θ is Euclidean valued – i.e. $\mathbb{D} = \mathbf{R}^m$ for some $m \in \mathbf{N}$, we have a simpler form of the lower bound in (3.10). This includes semiparametric and nonparametric models as well as parametric ones; see Examples 2.1 and 2.2.

Corollary 3.1 *Let $(\mathscr{X}_n, \mathcal{A}_n, \{P_{n,h} : h \in H\})$ be a sequence of statistical experiments, and $\hat{\theta}_n$ an estimator for the parameter $\theta : \{P_{n,h}\} \to \mathbb{D}_\phi \subset \mathbb{D}$ with $\mathbb{D} = \mathbf{R}^m$ for some $m \in \mathbf{N}$. Suppose that Assumptions 2.1, 2.2, 2.3, 3.1, 3.2 and 3.3 hold. If the efficiency bound $\Sigma_0 \equiv \langle \tilde{\theta}_0, \tilde{\theta}_0^\intercal \rangle$ is nonsingular, then it follows that*

$$\sup_{I \subset_f H} \liminf_{n \to \infty} \sup_{h \in I} E_{n,h}[\ell(r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(h))\})] \\ \geq \inf_{u \in \mathbf{R}^m} \sup_{c \in \mathbf{R}^m} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c))] \ . \tag{3.12}$$

The lower bound in (3.12) is a minimax optimization problem over $\mathbf{R}^m$; in particular, the supremum is taken over $\mathbf{R}^m$ instead of the tangent set. This simply follows from the facts that the support of $\mathbb{G}_0$ is $\overline{\theta'_0(H)}$ by Theorem 2.1 and that a nondegenerate Gaussian random variable in $\mathbf{R}^m$ has support $\mathbf{R}^m$. As a result, the construction of optimal plug-in estimators in Section 3.3 becomes much easier when θ is Euclidean valued.

3.2.1 Examples Revisited

In this section we explicitly derive the lower bound for each example introduced in Section 2.1.1. For simplicity of illustration, we confine our attention to the simplest i.i.d. setup. That is, we assume that the sample $X_1, \ldots, X_n$ is i.i.d. and distributed according to $P \in \mathcal{P}$, and we are interested in estimating $\phi(\theta)$.

Example 2.1 (Continued). Simple algebra reveals that ϕ'_θ is Lipschitz continuous. In order to compare with previous literature, consider the case when X is bivariate normal with covariance matrix $\sigma^2 I_2$, and take the squared loss function. As shown in Appendix B, the lower bounds is given by

$$\inf_{u \in \mathbf{R}^2} \sup_{c \in \mathbf{R}^2} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c))] \\ = \inf_{u \in \mathbf{R}^2} \sup_{c \in \mathbf{R}^2} E\left[\left(\max\{\mathbb{G}_0^{(1)} + u^{(1)} + c^{(1)}, \mathbb{G}_0^{(2)} + u^{(2)} + c^{(2)}\} - \max\{c^{(1)}, c^{(2)}\}\right)^2\right] = \sigma^2 \ ,$$

where $\mathbb{G}_0 \equiv (\mathbb{G}_0^{(1)}, \mathbb{G}_0^{(2)}) \sim N(0, \sigma^2 I_2)$, and the infimum is achieved when $u = (-\infty, 0)$ and $c = (-\infty, c^{(2)})$ with $c^{(2)} \in \mathbf{R}$ arbitrary. In fact, the lower bound can be also achieved at $u = 0$ and $c = 0$. We note that this lower bound is consistent with Song (2014a) and Blumenthal and Cohen (1968b). ■

Example 2.2 (Continued). In this case, it is also easy to see that ϕ'_θ is Lipschitz continuous. For the squared loss function, the lower bound at the point θ_0 with $\psi(\theta_0) = 0$ becomes

$$\inf_{u \in \mathbf{R}^2} \sup_{c \in \mathbf{R}^2} E[(\max\{\psi'_{\theta_0}(\mathbb{G}_0 + u + c), 0\} - \max\{\psi'_{\theta_0}(c), 0\})^2] \ ,$$

where $\mathbb{G}_0$ is the efficient Gaussian limit for estimating θ_0. ■

Example 2.3 (Continued). In this example, it can be shown that ϕ'_θ is Lipschitz continuous. For the loss function $\ell(z) = \|z\|_\infty$, the lower bound becomes

$$\inf_{u^{(1)},u^{(2)}\in\ell^\infty(\mathbf{R})} \sup_{h^{(1)},h^{(2)}\in H} \{E[\|(\mathbb{G}_0^{(1)} + u^{(1)})1_{B_1} + (\mathbb{G}_0^{(2)} + u^{(2)})1_{B_2} \\ + \max\{\mathbb{G}_0^{(1)} + u^{(1)} + h^{(1)}, \mathbb{G}_0^{(2)} + u^{(2)} + h^{(2)}\}1_{B_0} - \max\{h^{(1)}, h^{(2)}\}1_{B_0}\|_\infty]\} \ ,$$

where H consists of all bounded measurable real valued functions on $\mathbf{R}$ with $\int_{\mathbf{R}} h\, dP = 0$, and $(\mathbb{G}_0^{(1)}, \mathbb{G}_0^{(2)})$ is the efficient Gaussian limit in $\ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ for estimating $\theta_0 \equiv (F_1, F_2)$. ■

Example 2.4 (Continued). Since C_{θ_0} is closed and convex, ϕ'_{θ_0} or equivalently $\Pi_{C_{\theta_0}}$ is Lipschitz continuous (Zarantonello, 1971, p.241). If the loss function $\ell(\cdot) : L^2(\mathcal{T}) \to \mathbf{R}$ is $\ell(z) = \|z\|_{L^2}^2$, then the lower bound is finite and given by

$$\inf_{u\in L^2(\mathcal{T})} \sup_{h\in H} E[\|\Pi_{C_{\theta_0}}(\mathbb{G}_0 + u + \theta'_0(h)) - \Pi_{C_{\theta_0}}(\theta'_0(h))\|_{L^2}^2] \ ,$$

where $H \equiv \{(h_1, h_2) : h_1 \in H_1, h_2 \in H_2\}$ with[15]

$$H_1 \equiv \{h_1 : \mathscr{Z} \to \mathbf{R} : E[h_1(Z)] = 0\} \ , \\ H_2 \equiv \{h_2 : \mathscr{Y} \times \mathscr{Z} \to \mathbf{R} : E[h_2(Y, z)] = 0 \text{ for a.s. } z \in \mathscr{Z}\} \ ,$$

$\mathbb{G}_0$ is a zero mean Gaussian process in $L^2(\mathcal{T})$ with covariance function $\text{Cov}(\tau_1, \tau_2) \equiv J(\tau_1)^{-1}\Gamma(\tau_1, \tau_2)J(\tau_2)^{-1}$ in which for $f_Y(y|Z)$ the density of Y conditional on Z,

$$J(\tau) \equiv c' E[f_Y(Z'\beta(\tau)|Z)ZZ'] \ , \forall\, \tau \in \mathcal{T} \ , \\ \Gamma(\tau_1, \tau_2) \equiv E[(\tau_1 - 1\{Y \le Z'\beta(\tau_1)\})(\tau_2 - 1\{Y \le Z'\beta(\tau_2)\})ZZ'] \ , \forall\, \tau_1, \tau_2 \in \mathcal{T} \ ,$$

and,

$$\theta'_0(h)(\tau) \equiv -J(\tau)^{-1} \int c'z1\{y \le z'\beta(\tau)\}h_1(y, z)P(dy, dz) \\ - J(\tau)^{-1} \int c'z(1\{y \le z'\beta(\tau)\} - \tau)h_2(z)P(dy, dz) \ .$$

For a detailed discussion on the efficient estimation of θ, see Lee (2009, Theorem 3.1). ■

3.3 Attainability via Construction

Having established the lower bounds as in Theorem 3.1 and Corollary 3.1, we now proceed to show the attainability of the bounds by developing a general procedure of constructing optimal plug-in estimators. The lower bounds in (3.10) and (3.12) suggest that an optimal plug-in estimator is of the form $\phi(\hat{\theta}_n + \hat{u}_n/r_n)$ where $\hat{u}_n$ is an estimator of the optimal noise term in Theorem 2.1 – i.e. $\hat{u}_n$ should be an estimator of the minimizer(s) in the lower bounds. We deal with infinite dimensional $\mathbb{D}$ first in order to accommodate Examples 2.3 and 2.4, and then specialize to Euclidean $\mathbb{D}$.

[15] see Severini and Tripathi (2001). Technically, H here is not the tangent set; however, every element in the tangent set can be written as a unique decomposition involving some pair in H. This shouldn't bother us since tangent set *per se* is not of our interest.

Recall from Theorem 3.1 that the lower bound for the local asymptotic minimax risk is given by

$$\inf_{u\in\mathbb{D}} \sup_{h\in H} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h))\right)\right] \ . \tag{3.13}$$

If the objective function in (3.13) were known, we would pick the optimal correction term by solving a minimax optimization problem. However, this is not the case since there are four unknown objects here: the law of the efficient Gaussian component $\mathbb{G}_0$, the derivatives ϕ'_{θ_0} and θ'_0, and the space H. We thus work with the sample analog of (3.13) by replacing $\mathbb{G}_0$, ϕ'_{θ_0}, θ'_0, and H with their sample counterparts.

We shall assume that the law of $\mathbb{G}_0$ can be estimated by bootstrap or simulation. Specifically, let $\hat{\theta}_n$ be an efficient estimator of θ, and $\hat{\theta}^*_n$ a bootstrapped version of it – i.e. $\hat{\theta}^*_n$ is a function mapping the data $\{X_i\}_{i=1}^n$ and random weights $\{W_i\}$ that are independent of $\{X_i\}$ into the domain $\mathbb{D}_\phi$ of ϕ. This abstract definition suffices for encompassing the nonparametric, Bayesian, block, score, and weighted bootstrap as special cases. The hope is then that the limiting law of $r_n\{\hat{\theta}_n - \theta_0\}$ can be consistently estimated by the (finite sample) law of $r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}$, which necessitates a metric that measures distances between probability measures. Since the law $\mathbb{G}_0$ is tight and hence separable, we may employ the bounded Lipschitz metric d_{BL} introduced by Dudley (1966, 1968): for two Borel probability measures L_1 and L_2 on $\mathbb{D}$, define

$$d_{\text{BL}}(L_1, L_2) \equiv \sup_{f\in \text{BL}_1(\mathbb{D})} \left| \int f\, dL_1 - \int f\, dL_2 \right| ,$$

where recall that $\text{BL}_1(\mathbb{D})$ is the space of bounded and Lipschitz continuous functions as defined in (2.3). We may now measure the distance between the law of $\hat{\mathbb{G}}^*_n \equiv r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}$ conditional on $\{X_i\}$ and the limiting law $\mathbb{G}_0$ of $r_n\{\hat{\theta}_n - \theta_0\}$ by

$$d_{\text{BL}}(\hat{\mathbb{G}}^*_n, \mathbb{G}_0) = \sup_{f\in \text{BL}_1(\mathbb{D})} \left| E[f(r_n\{\hat{\theta}^*_n - \hat{\theta}_n\})|\{X_i\}] - E[f(\mathbb{G}_0)] \right| . \tag{3.14}$$

Employing the distribution of $r_n\{\hat{\theta}^*_n - \hat{\theta}_n\}$ conditional on the data to approximate the distribution of $\mathbb{G}_0$ is then asymptotically justified if their distance, equivalently (3.14), converges in probability to zero.

The estimation of θ'_0 can be done by analogy principle since the derivative θ'_0 typically takes the form $\theta'_0 \equiv \theta'_0(P)$, that is, we may estimate θ'_0 by $\hat{\theta}'_n = \theta'_0(\mathbb{P}_n)$ with $\mathbb{P}_n$ the empirical measure. Estimation of the derivative ϕ'_{θ_0} is trickier. In this regard, we impose sufficient conditions so as to meet Assumption 3.3 in Fang and Santos (2014). The following assumption formalizes our discussion so far.

Assumption 3.5 *(i)* $\hat{\mathbb{G}}^*_n : \{X_i, W_i\}_{i=1}^n \to \mathbb{D}_\phi$ *with* $\{W_i\}$ *independent of* $\{X_i\}$ *satisfies* $\sup_{f\in \text{BL}_1(\mathbb{D})} |E[f(\hat{\mathbb{G}}^*_n)|\{X_i\}] - E[f(\mathbb{G}_0)]| = o_p(1)$ *under* $\{P_{n,0}\}$.

(ii) $\hat{\theta}'_n : H \to \mathbb{D}$ *depends on* $\{X_i\}$ *and satisfies* $\|\hat{\theta}'_n(\hat{h}_n) - \theta'_0(h)\|_{\mathbb{D}} \xrightarrow{p} 0$ *under* $\{P_{n,0}\}$ *whenever* $\|\hat{h}_n - h\|_H \xrightarrow{p} 0$ *under* $\{P_{n,0}\}$ *with* $\hat{h}_n : \{X_i\} \to H$.

(iii) $\hat{\phi}'_n : \mathbb{D} \to \mathbb{E}$ *depends on* $\{X_i\}$ *satisfying (a) for any* $z \in \mathbb{D}$, $\hat{\phi}'_n(z)$ *is consistent for* $\phi'_{\theta_0}(z)$ *– i.e.* $\|\hat{\phi}'_n(z) - \phi'_{\theta_0}(z)\|_{\mathbb{E}} \xrightarrow{p} 0$ *under* $\{P_{n,0}\}$*; and (b) there is some deterministic constant* $C_{\hat{\phi}'}$ *such that* $\|\hat{\phi}'_n(z_1) - \hat{\phi}'_n(z_2)\|_{\mathbb{E}} \leq C_{\hat{\phi}'}\|z_1 - z_2\|_{\mathbb{D}}$ *outer almost surely for all* $z_1, z_2 \in \mathbb{D}$.

Assumption 3.5(i) is simply a bootstrap consistency condition on $\hat{\mathbb{G}}_n^*$ for the target law of $\mathbb{G}_0$, including Song (2014a)'s simulation method as a special case. Assumption 3.5(ii) imposes a weak consistency condition on the estimator $\hat{\theta}_n$. One might require $\hat{\theta}'_n$ be consistent in the sense that $\|\hat{\theta}'_n - \theta'_0\|_{op} \xrightarrow{p} 0$ where $\|\cdot\|_{op}$ is the operator norm. However, such an assumption is too restrictive for a Glivenko-Cantelli argument to hold since the operator norm is a supremum taken over all $h \in H$ with $\|h\|_H \leq 1$. The pointwise consistency condition on $\hat{\phi}'_n$ in Assumption 3.5(iii)-(a) is a minimal requirement, while Assumption 3.5(iii)-(b) imposes Lipschitz continuity on $\hat{\phi}'_n$, a condition inherited from ϕ'_{θ_0} as in Assumption 3.4. Assumptions 3.5(iii)-(a) and -(b) together imply that $\hat{\phi}'_n$ converges in probability to ϕ_{θ_0} uniformly over all δ-enlargement of compact sets in $\mathbb{D}$, a condition that has been employed in Fang and Santos (2014) to construct a valid inference procedure for the parameter $\phi(\theta)$.

We next deal with approximating the spaces H and $\mathbb{D}$ as needed to construct an analog to the bound (3.13). To understand the unknown nature of H, consider the i.i.d. setup in which case $H \equiv \dot{\mathcal{P}}^0$ where $\dot{\mathcal{P}}^0$ is the tangent set as defined in Remark 2.2. In these settings, it is common that $\dot{\mathcal{P}}^0$ is equal to the largest possible tangent set $L_0^2(P) \equiv \{h \in L^2(P) : \int h dP = 0\}$, which depends on the unknown probability measure P. It is worth noting that $L_0^2(P)$ can be viewed as the projection of $L^2(P)$ onto the complement of the subspace of constant functions. In fact, this projection nature of $\dot{\mathcal{P}}^0$ is prevalent in efficient estimation (Bickel et al., 1993), an insight helpful to the estimation of H.

Since both H and $\mathbb{D}$ are infinite dimensional, we need to approximate H and $\mathbb{D}$ by sequences of sieve spaces, which typically consist of compact subsets or finite dimensional subspaces that grow dense in H and $\mathbb{D}$. Consider the space H first. If we have a "basis" $\{g_m\}$ for $\dot{\mathcal{P}}^0$, then we may approximate H by finite dimensional subspaces constructed from $\{g_m\}$. For example, the space $C_c^0(\mathbf{R}^{d_x})$ of mean zero continuous functions on $\mathbf{R}^{d_x}$ with compact support is dense in $L_0^2(P)$; by the Stone-Weierstrass theorem, the set of polynomial functions are in turn dense in $C_c^0(\mathbf{R}^{d_x})$. Thus, following Chamberlain (1987) who approximates the efficiency bound in models defined by conditional moment restrictions based on polynomials, we may take the polynomials, properly projected or truncated, as a complete sequence in $\overline{H}$. As for the space $\mathbb{D}$ over which the infimum is taken, we may employ linear sieves as approximation. These being said, we assume the following:

Assumption 3.6 *(i) $\{g_m\}_{m=1}^{\infty} \subset H$ is complete in the sense that for each $h \in H$ and $\epsilon > 0$, there exists $\alpha_1, \ldots, \alpha_m$ such that $\|h - \sum_{j=1}^m \alpha_j g_j\|_H < \epsilon$; (ii) for each $m \in \mathbf{N}$, $\hat{g}_m : \{X_i\} \to H$ satisfies $\|\hat{g}_m - g_m\|_H \xrightarrow{p} 0$ under $\{P_{n,0}\}$; (iii) $\{\psi_k\}_{k=1}^{\infty} \subset \mathbb{D}$ is complete.*

Assumption 3.6(i) formalizes the approximation property of $\{g_m\}$, in a way like the Schauder basis except that the representation coefficients α_j might not be unique, while Assumption 3.6(iii) is a similar approximation condition imposed on $\{\psi_k\}$. Assumption 3.6(ii) requires that $\{g_m\}$ be estimated by a sequence $\{\hat{g}_m\}$ of random variables to accommodate the unknown nature of H.

Given the availability of complete sequences $\{g_m\}$ and $\{\psi_k\}$ in H and $\mathbb{D}$ respectively, we may approximate the lower bound (3.13) by

$$\min_{v \in K_{\tau_k}^k} \max_{c \in K_{\lambda_m}^m} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(g^m)^\intercal c) - \phi'_{\theta_0}(\theta'_0(g^m)^\intercal c))] \ , \tag{3.15}$$

where $K^k_{\tau_k}$ and $K^m_{\lambda_m}$ are balls in $\mathbf{R}^k$ and $\mathbf{R}^m$ respectively as defined in the beginning of Section 2.1, $\{\lambda_m\}$ and $\{\tau_k\}$ are sequences that diverge to infinity as $m, k \to \infty$ respectively, and $\theta_0'(g^m) \equiv (\theta_0'(g_1), \ldots, \theta_0'(g_m))^\intercal$. Heuristically, (3.15) is the bound for the parametric submodel whose tangent set is $\{c^\intercal g^m : c \in K^m_{\lambda_m}\}$ and noise term u is restricted to be bounded in norm by τ_k. As the approximation indices m, k increase to infinity, (3.15) converges to the lower bound (3.13). With g^m, $\mathbb{G}_0$, θ_0' and ϕ'_{θ_0} in (3.15) replaced by the corresponding estimates $\{\hat{g}_m\}$, $\hat{\mathbb{G}}^*_n$, $\hat{\theta}'_n$ and $\hat{\phi}'_n$, the bound (3.15) can in turn be estimated by

$$\min_{v \in K^k_{\tau_k}} \max_{c \in K^m_{\lambda_m}} E[\ell(\hat{\phi}'_n(\hat{\mathbb{G}}^*_n + (\psi^k)^\intercal v + \hat{\theta}'_n(\hat{g}^m)^\intercal c) - \hat{\phi}'_n(\hat{\theta}'_n(\hat{g}^m)^\intercal c))|\{X_i\}] \ , \tag{3.16}$$

where $\hat{\theta}'_n(\hat{g}^m) \equiv (\hat{\theta}'_n(\hat{g}_1), \ldots, \hat{\theta}'_n(\hat{g}_m))^\intercal$, and the expectation is evaluated with respect to the bootstrap weights $\{W_i\}_{i=1}^n$ holding $\{X_i\}_{i=1}^n$ fixed. For notational simplicity, define

$$\hat{B}_m(v) \equiv \max_{c \in K^m_{\lambda_m}} E[\ell(\hat{\phi}'_n(\hat{\mathbb{G}}^*_n + (\psi^k)^\intercal v + \hat{\theta}'_n(\hat{g}^m)^\intercal c) - \hat{\phi}'_n(\hat{\theta}'_n(\hat{g}^m)^\intercal c))|\{X_i\}] \ ,$$
$$\hat{\Psi}_{k,m} \equiv \{v \in K^k_{\tau_k} : \hat{B}_m(v) \leq \min_{v' \in K^k_{\tau_k}} \hat{B}_m(v') + \epsilon_n\} \ ,$$

where $\epsilon_n = o_p(1)$ as $n \to \infty$. Here, $\hat{\Psi}_{k,m}$ is the set of minimizers for the sample analog approximating problem (3.16), allowing negligible computational error ϵ_n that tends to zero in probability.

We are now ready to construct the optimal plug-in estimators. For any $\hat{v}_{n,k,m} \in \hat{\Psi}_{k,m}$, we consider estimating $\phi(\theta_n(h))$ by

$$\phi(\hat{\theta}_n + \frac{\hat{u}_{n,k,m}}{r_n}) \ , \quad \hat{u}_{n,k,m} \equiv (\psi^k)^\intercal \hat{v}_{n,k,m} \ , \tag{3.17}$$

where $\hat{\theta}_n$ is an efficient estimator of θ – i.e. it satisfies

Assumption 3.7 *$\{\hat{\theta}_n\}$ is an efficient estimator of θ – i.e. for each $h \in H$,*

$$r_n\{\hat{\theta}_n - \theta_n(h)\} \overset{L_{n,h}}{\to} \mathbb{G}_0 \text{ in } \mathbb{D} \ ,$$

where $\mathbb{G}_0$ is the efficient Gaussian random variable as in Theorem 2.1.

Our first construction result shows that the plug-in estimator (3.17) attains the local asymptotic minimax lower bound (3.13).

Theorem 3.2 *Suppose that Assumptions 2.1, 2.2, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, and 3.7 hold. Let $\{\lambda_m\}$ and $\{\tau_k\}$ be sequences that diverge to infinity as $m, k \to \infty$ respectively. If $\hat{v}_{n,k,m} \in \hat{\Psi}_{k,m}$, then*

$$\limsup_{k \to \infty} \limsup_{m \to \infty} \sup_{I \subset_f H} \limsup_{n \to \infty} \sup_{h \in I} E_{n,h}\big[\ell\big(r_n\big(\phi(\hat{\theta}_n + \frac{\hat{u}_{n,k,m}}{r_n}) - \phi(\theta_n(h))\big)\big)\big]$$
$$\leq \inf_{u \in \mathbb{D}} \sup_{h \in H} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta_0'(h)) - \phi'_{\theta_0}(\theta_0'(h))\right)\right] \ , \tag{3.18}$$

where $\hat{u}_{n,k,m} \equiv (\psi^k)^\intercal \hat{v}_{n,k,m}$.

We note that, though unpleasant, the first two lim sup's over k and m are necessary in general and more importantly are taken after letting $n \to \infty$. The reason is that minimizers in $\hat{\Psi}_{k,m}$ would possibly diverge to "infinity" as the search ranges $K^k_{\tau_k}$ and $K^m_{\lambda_m}$ grow to the whole (noncompact) spaces, rendering the Delta method inapplicable under just Hadamard directional differentiability. Nonetheless, by restricting u to be in a compact set $\mathbb{D}_u \subset \mathbb{D}$, for example a class of smooth functions, we are able to remove the first lim sup; see Section 3.3.1.

The general construction of optimal plug-in estimators for infinite dimensional $\mathbb{D}$ is intrinsically complicated. When $\mathbb{D}$ is Euclidean – i.e. $\mathbb{D} = \mathbf{R}^m$ for some $m \in \mathbf{N}$, the computation greatly simplifies. Recall that by Corollary 3.1, the lower bound in this case is given by

$$\inf_{u \in \mathbf{R}^m} \sup_{c \in \mathbf{R}^m} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c)\right)\right] \ . \tag{3.19}$$

Comparing (3.19) with (3.13), it is clear that we can dispense with the computation burden of estimating H and θ'_0. Instead we now only have to estimate the directional derivative ϕ'_{θ_0} and the law of $\mathbb{G}_0$. Following the same idea as before, we therefore define

$$\begin{aligned} \hat{B}_\lambda(u) &\equiv \max_{c \in K^m_\lambda} E[\ell(\hat{\phi}'_n(\hat{\mathbb{G}}^*_n + u + c) - \hat{\phi}'_n(c))|\{X_i\}] \ , \\ \hat{\Psi}_{\tau,\lambda} &\equiv \{u \in K^m_\tau : \hat{B}_\lambda(u) \leq \min_{u' \in K^m_\tau} \hat{B}_\lambda(u') + \epsilon_n\} \ , \end{aligned}$$

where $\epsilon_n = o_p(1)$ as $n \to \infty$. As expected, if we pick $\hat{u}_{n,\tau,\lambda} \in \hat{\Psi}_{\tau,\lambda}$, then

$$\phi\big(\hat{\theta}_n + \frac{\hat{u}_{n,\tau,\lambda}}{r_n}\big) \tag{3.20}$$

will be an optimal plug-in estimator, as confirmed by the following theorem.

Theorem 3.3 *Let $\mathbb{D} = \mathbf{R}^m$ for some $m \in \mathbf{N}$ and $\Sigma_0 \equiv \langle \tilde{\theta}_0, \tilde{\theta}_0^\intercal \rangle$ be nonsingular. Suppose that Assumptions 2.1, 2.2, 3.1, 3.2, 3.3, 3.4, 3.5(i)(iii), and 3.7 hold. Then*

$$\begin{aligned} \limsup_{\tau \to \infty} \limsup_{\lambda \to \infty} \sup_{I \subset_f H} \limsup_{n \to \infty} \sup_{h \in I} & E_{n,h}\big[\ell\big(r_n\big(\phi(\hat{\theta}_n + \frac{\hat{u}_{n,\tau,\lambda}}{r_n}) - \phi(\theta_n(h))\big)\big)\big] \\ & \leq \inf_{u \in \mathbf{R}^m} \sup_{c \in \mathbf{R}^m} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c)\right)\right] \ . \end{aligned} \tag{3.21}$$

It is worth noting that the optimal plug-in estimators (3.17) and (3.20) depend, through the correction terms $\hat{u}_{n,m,k}$ and $\hat{u}_{n,\tau,\lambda}$ respectively, on the choice of the loss function ℓ, which in turn hinges on the nature of the problem at hand and practitioners' risk preference.

3.3.1 Smoothed Optimal Plug-in Estimators

By letting $k, m \to \infty$ and $\tau, \lambda \to \infty$ after n tends to infinity in the lower bounds, one essentially confines the minimizers $\hat{u}_{n,m,k}$ and $\hat{u}_{n,\tau,\lambda}$ to compact subsets. We may alternatively start with compact (possibly infinite dimensional) spaces and base our analysis therein.

In the literature of nonparametric and semi-(non)parametric methods, compactness can be obtained by attaching an appropriate norm different from the one that defines the space under consideration (Gallant and Nychka, 1987). For detailed discussions we refer the readers to Gallant and Nychka (1987), Newey and Powell (2003) and Santos (2012). We instead impose the following high level conditions.

Assumption 3.8 *(i) $\mathbb{D}_u \subset \mathbb{D}$ is compact; (ii) $\{\mathbb{D}_k\}_{k=1}^{\infty}$ with $\mathbb{D}_k \subset \mathbb{D}_u$ for each $k \in \mathbf{N}$ is a sequence of compact sieves satisfying for any $u \in \mathbb{D}_u$, there exists $u_k \in \mathbb{D}_k$ such that $\|u_k - u\|_{\mathbb{D}} \to 0$ as $k \to \infty$.*

Suppose that we are interested in the following restricted version of lower bound:

$$\min_{u \in \mathbb{D}_u} \sup_{h \in H} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h))\right)\right] , \tag{3.22}$$

which is equal to the bound (3.13) if the infimum in the latter is attained in $\mathbb{D}_u$. In turn, (3.22) can be approximated by

$$\min_{u \in \mathbb{D}_u} \max_{c \in K^m_{\lambda_m}} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(g^m)^\intercal c) - \phi'_{\theta_0}(\theta'_0(g^m)^\intercal c))] , \tag{3.23}$$

where $\lambda_m \to \infty$ as $m \to \infty$. Replacing g^m, $\mathbb{G}_0$, θ'_0 and ϕ'_{θ_0} in (3.23) by their corresponding estimates $\{\hat{g}_m\}$, $\hat{\mathbb{G}}^*_n$, $\hat{\theta}'_n$ and $\hat{\phi}'_n$, and approximating $\mathbb{D}_u$ by the sequence of compact sieves $\{\mathbb{D}_k\}$, we may in turn estimate the bound (3.23) by considering

$$\begin{aligned}
\hat{B}_m(u) &\equiv \max_{c \in K^m_{\lambda_m}} E[\ell(\hat{\phi}'_n(\hat{\mathbb{G}}^*_n + u + \hat{\theta}'_n(\hat{g}^m)^\intercal c) - \hat{\phi}'_n(\hat{\theta}'_n(\hat{g}^m)^\intercal c))|\{X_i\}] , \\
\hat{\Psi}_m &\equiv \{u \in \mathbb{D}_{k_n} : \hat{B}_m(u) \leq \min_{u' \in \mathbb{D}_{k_n}} \hat{B}_m(u') + \epsilon_n\} ,
\end{aligned}$$

where $\epsilon_n = o_p(1)$ as $n \to \infty$. Notice that the set $\hat{\Psi}_m$ of minimizers of $\hat{B}_m(u)$ is obtained on the approximating space $\mathbb{D}_{k_n}$, though we have suppressed the dependence of $\hat{\Psi}_m$ on n for notational simplicity.

Now take arbitrary $\hat{u}_{n,m} \in \hat{\Psi}_m$ and define the plug-in estimator

$$\phi(\hat{\theta}_n + \frac{\hat{u}_{n,m}}{r_n}) . \tag{3.24}$$

Optimality of (3.24) in the sense of local asymptotic minimaxity is confirmed as follows.

Theorem 3.4 *Suppose that Assumptions 2.1, 2.2, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6(i)(ii), 3.7, and 3.8 hold. Let $\hat{u}_{n,m} \in \hat{\Psi}_m$. If $\lambda_m, k_n \to \infty$ as $m, n \to \infty$ respectively, then*

$$\begin{aligned}
&\limsup_{m \to \infty} \sup_{I \subset_f H} \limsup_{n \to \infty} \sup_{h \in I} E_{n,h}\left[\ell\big(r_n\big(\phi(\hat{\theta}_n + \frac{\hat{u}_{n,m}}{r_n}) - \phi(\theta_n(h))\big)\big)\right] \\
&\qquad \leq \inf_{u \in \mathbb{D}_u} \sup_{h \in H} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h))\right)\right] .
\end{aligned} \tag{3.25}$$

We note that similar as the sieve approximation for $\mathbb{D}$, one may also consider construction based on a general sequence of compact sieves of H. While one might have different tastes on the choice of compact sieves for $\mathbb{D}$ – for instance, one might choose different degrees of smoothness which in turn directly affects the smoothness of the correction term $\hat{u}_n$, approximation for H is purely for computational purposes and has more indirect effect on $\hat{u}_n$. We thus skip the general approximation for H here.

3.3.2 Examples Revisited

We now turn to Examples 2.1-2.4. For the sake of brevity, we omit the bootstrap procedure, and instead focus on verifying Assumptions 3.5(ii)(iii), 3.6, and 3.7. For Examples 2.1 and 2.2, there is no need to estimate H and θ_0'; see Corollary 3.1.

Example 2.1 (Continued). The sample mean $\overline{X}_n$ serves as an efficient estimator of θ. Denote $\hat{j}^* = \arg\max_{j\in\{1,2\}} \bar{X}^{(j)}$ and pick $t_n \uparrow \infty$ satisfying $t_n/\sqrt{n} \downarrow 0$. Define

$$\hat{\phi}_n'(z) = \begin{cases} z^{(\hat{j}^*)} & \text{if } |\bar{X}^{(1)} - \bar{X}^{(2)}| > t_n \\ \max\{z^{(1)}, z^{(2)}\} & \text{if } |\bar{X}^{(1)} - \bar{X}^{(2)}| \le t_n \end{cases} . \tag{3.26}$$

Then it is straightforward to verify that $\hat{\phi}_n'$ is Lipschitz continuous and pointwise consistent. ■

Example 2.2 (Continued). The efficient estimation of θ_0 in this example can be conducted in the conditional moment restriction framework (Newey, 1993). Then we may estimate ϕ_{θ_0}' by

$$\hat{\phi}_n'(z) = \psi_{\hat{\theta}_n}'(z) 1\{\psi(\hat{\theta}_n) > t_n\} + \max\{\psi_{\hat{\theta}_n}'(z), 0\} 1\{|\psi(\hat{\theta}_n)| \le t_n\} ,$$

where $\hat{\theta}_n$ is an efficient estimator of θ_0, and t_n is a sequence specified as in Example 2.1. ■

Example 2.3 (Continued). Let $\hat{F}_1$ and $\hat{F}_2$ be the empirical CDFs of F_1 and F_2 respectively. It is known that empirical CDFs $\hat{F}_1$ and $\hat{F}_2$ are efficient in estimating F_1 and F_2 respectively (van der Vaart and Wellner, 1996), and hence $(\hat{F}_1, \hat{F}_2)$ is efficient in estimating (F_1, F_2) in the product space $\ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ (van der Vaart, 1991b). The form of the derivative ϕ_{θ_0}' as in (3.5) suggests a natural estimator for it. Define $\hat{B}_1 \equiv \{x \in \mathbf{R} : \hat{F}_{1n}(x) - \hat{F}_{2n}(x) > t_n\}$, $\hat{B}_2 \equiv \{x \in \mathbf{R} : \hat{F}_{2n}(x) - \hat{F}_{1n}(x) > t_n\}$, and $\hat{B}_0 \equiv \{x \in \mathbf{R} : |\hat{F}_{1n}(x) - \hat{F}_{2n}(x)| \le t_n\}$ where t_n is again as in Example 2.1. Let $\hat{\phi}_n' : \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R}) \to \ell^\infty(\mathbf{R})$ be defined by

$$\hat{\phi}_n'(z) = z^{(1)} 1_{\hat{B}_1} + z^{(2)} 1_{\hat{B}_2} + \max\{z^{(1)}, z^{(2)}\} 1_{\hat{B}_0} .$$

Then $\hat{\phi}_n'$ is Lipschitz continuous and pointwise consistent.

In this example, we have to estimate θ_0' and a basis $\{g_j\}$ of $L_0^2(P)$, as well as the derivative ϕ_{θ_0}' and the law of $\mathbb{G}_0$. By Section 3.11.1 in van der Vaart and Wellner (1996), for each $h \equiv (h_1, h_2) \in H \times H$ with H being the set of bounded measurable functions on $\mathbf{R}$,

$$\theta_0'(h)(v) = \big(\int_{-\infty}^v h_1(t)\, P_1(dt), \int_{-\infty}^v h_2(t)\, P_2(dt) \big) .$$

Thus, we may take the following estimator of θ_0':

$$\hat{\theta}_n'(h)(v) = \big(\int_{-\infty}^v h_1(t)\, \mathbb{P}_{1n}(dt), \int_{-\infty}^v h_2(t)\, \mathbb{P}_{2n}(dt) \big) .$$

As to Assumption 3.6(ii), if $\{g_m^{(i)}\}$ is complete in $L^2(P_i)$ with $i = 1, 2$, then we may take

$$g_1^{(1)}(v) - \frac{1}{n}\sum_{i=1}^n g_1^{(1)}(B_{1i}) \ , \ g_2^{(1)}(v) - \frac{1}{n}\sum_{i=1}^n g_2^{(1)}(B_{1i}) \ , \dots \ ,$$

$$g_1^{(2)}(v) - \frac{1}{n}\sum_{i=1}^n g_1^{(2)}(B_{2i}) \ , \ g_2^{(2)}(v) - \frac{1}{n}\sum_{i=1}^n g_2^{(2)}(B_{2i}) \ , \dots \ ,$$

where $\{B_{1i}\}_{i=1}^n$ and $\{B_{2i}\}_{i=1}^n$ are bids from auctions 1 and 2 respectively; see Lemma B.2.[16] In this example, since functions in $\ell^\infty(\mathbf{R})$ can be rather irregular, one might want to follow the compact version of construction, for instance, let $\mathbb{D}_u$ be a class of smooth $\mathbf{R}^2$-valued functions. For concrete constructions, see Gallant and Nychka (1987), Newey and Powell (2003), and Santos (2012). ■

Example 2.4 (Continued). Since $\beta(\cdot) : \mathcal{T} \to \mathbf{R}$ can be efficiently estimated by the quantile regression process $\hat{\beta}_n(\cdot)$, we thus conclude that $\hat{\theta}_n \equiv c'\hat{\beta}_n(\cdot)$ is efficient in estimating θ_0 (van der Vaart, 1991b). As to estimation of the derivative ϕ'_{θ_0}, we follow the approach pursued by Hong and Li (2014) and propose the following estimator:

$$\hat{\phi}'_n(z) \equiv t_n^{-1}\{\Pi_\Lambda(\hat{\theta}_n + t_n z) - \Pi_\Lambda(\hat{\theta}_n)\} ,$$

where t_n satisfies $t_n \to 0$ and $t_n\sqrt{n} \to \infty$ as $n \to \infty$.[17] The derivative θ'_0 can be estimated as follows:

$$\begin{aligned}\hat{\theta}'_n(h) \equiv -\hat{J}(\tau)^{-1} &\int c'z1\{y \le z'\hat{\beta}(\tau)\}h_1(y,z)\mathbb{P}_n(dy,dz) \\ &- \hat{J}(\tau)^{-1}\int c'z(1\{y \le z'\hat{\beta}(\tau)\} - \tau)h_2(z)\mathbb{P}_n(dy,dz) ,\end{aligned}$$

where $\hat{J}(\tau)$ is constructed as in Angrist et al. (2006):

$$\hat{J}(\tau) \equiv \frac{1}{2n\kappa_n}\sum_{i=1}^n 1\{|Y_i - Z_i'\hat{\beta}(\tau)| \le \kappa_n\}Z_iZ_i' ,$$

where κ_n satisfies $\kappa_n \to 0$ and $\kappa_n^2 n \to \infty$. A complete sequence in H_1 can be estimated similarly as in Example 2.3. As to H_2, if $\{g_j(y,z)\}$ is complete in $L^2(\mathscr{Y} \times \mathscr{Z})$, then we may take

$$g_1(y,z) - \frac{1}{n}\sum_{i=1}^n g_1(Y_i,z) , g_2(y,z) - \frac{1}{n}\sum_{i=1}^n g_2(Y_i,z) , \dots .$$

A complete sequence $\{\psi_k\}$ in $L^2(\mathcal{T})$ can be a sequence of polynomials, while the compact space $\mathbb{D}_u$ can be chosen to be a class of smooth functions in $L^2(\mathcal{T})$ as in Example 2.3. ■

4 Empirical Application

In this section, we apply the theory developed in previous sections to the estimation of the effect of Vietnam veteran status on the quantiles of civilian earnings (Angrist, 1990). Since certain types of men are more likely to service in the military, making the veteran status endogenous, a conventional quantile regression method is inappropriate to recover the casual relationship. Following Angrist (1990), we employ the Vietnam draft lottery eligibility indicator as an instrument for veteran status. In particular, we apply the instrumental quantile regression framework developed by Chernozhukov and Hansen (2005, 2006) to the Current Population Survey data set as in Chernozhukov et al.

[16] For example, if P_i satisfies $\int e^{M|v|} P_i(dv) < \infty$ for all $M \in (0,\infty)$, then $\{1, v, v^2, \dots\}$ is complete in $L^2(P_i)$.

[17] Song (2014a,b) essentially took the same approach.

(2010), which consists of four variables: annual labor real earnings, weakly real wage, veteran status indicator with value 1 for veterans, and Vietnam draft lottery eligibility indicator as an instrument with value 1 for eligible men. As in Chernozhukov et al. (2010), we focus on the annual labor earnings throughout.

Let Y denote the annual labor real earnings, D the veteran status, and Z the Vietnam draft lottery eligibility. Under instrument independence and rank similarity, Chernozhukov and Hansen (2005) showed that the quantile regression coefficients $\beta(\tau)$ for veterans can be identified by the following conditional moment restriction:

$$E[(\tau - 1\{Y \leq \beta(\tau)D\})|Z] = 0 \text{ a.s.}, \forall \tau \in (0,1) \ , \tag{4.1}$$

much like the counterpart in mean regression models. Chernozhukov and Hansen (2006) developed the instrumental variable quantile regression based on restriction (4.1), which can be viewed as a quantile regression analog of two stage least squares.

Unfortunately, since $\beta(\tau)$ is estimated pointwise, there is in general no guarantee that the quantile function $\hat{\beta}(\cdot)$ is monotonically increasing. To circumvent the non-monotonicity when estimating the structural quantile functions of earnings, we therefore employ the metric projection operator introduced in Example 2.4. In estimating the correction terms, we take polynomials as basis functions for $\mathbb{D} \equiv L^2(\mathcal{T})$ and H, and set $m = 4, k = 3$. The quantile index set $\mathcal{T}$ is taken to be the grid on $[0.25, 0.75]$ with increment 0.001, while the number of bootstrap repetitions is set to be two hundred. As for the estimation of the Hadamard directional derivative, we follow the same approach as in Example 2.4 and set $t_n = n^{-1/3}$. The correction terms are estimated relative to the L^2 loss function.

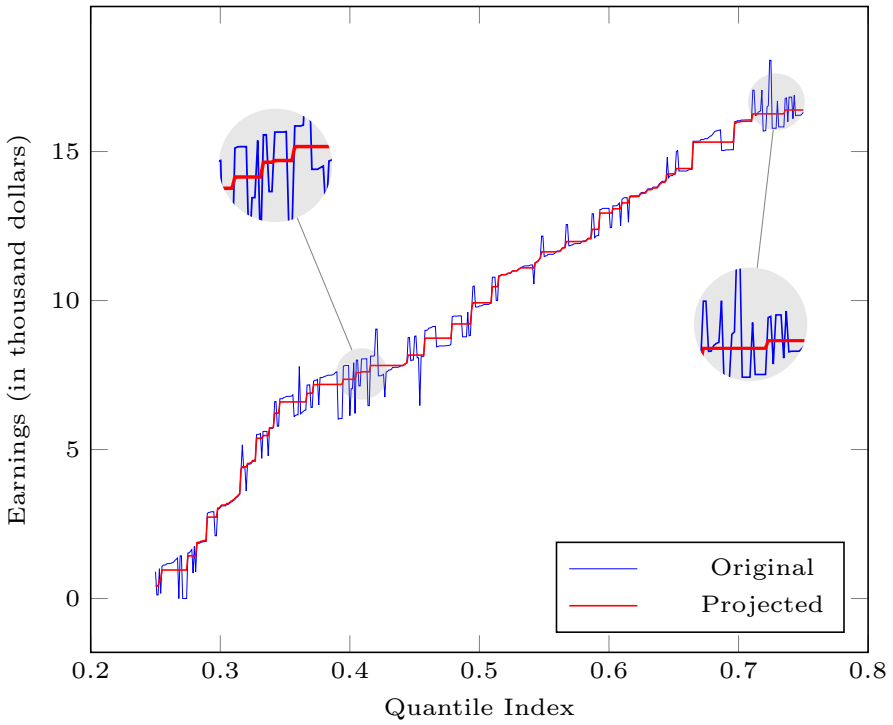

Figure 1 Structural Quantile Functions of Earnings for Veterans

In Figures 1 and 2 we show the structural quantile functions of earnings for veterans and non-veterans respectively, as well as their optimal projected counterparts with

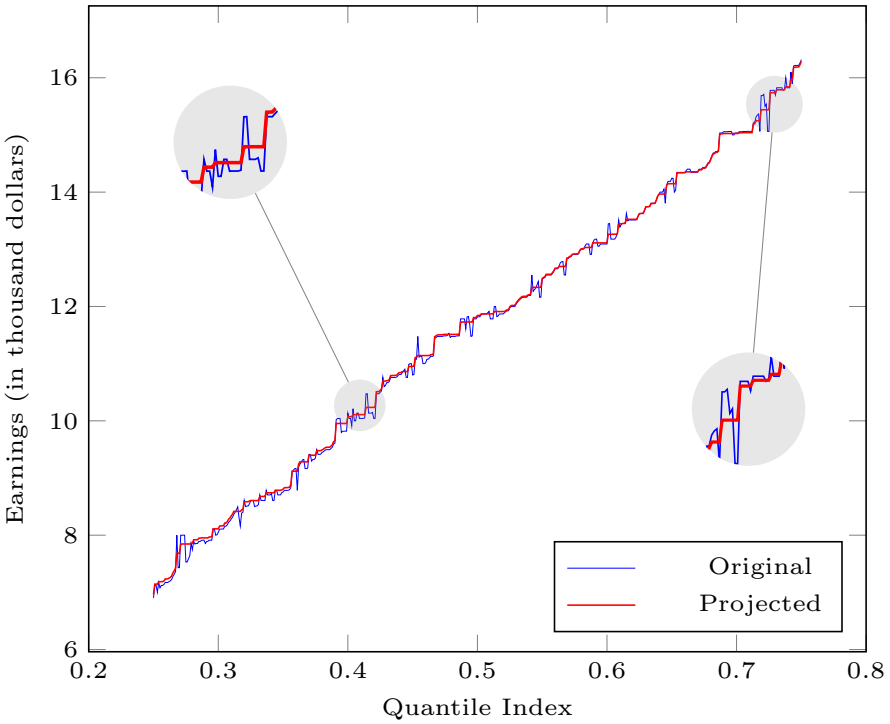

Figure 2 Structural Quantile Functions of Earnings for Non-Veterans

correction terms. In both figures, the original quantile functions exhibit obvious non-monotonicity at certain regions, especially for veterans. The projected counterparts are by construction monotone and optimal in terms of local asymptotic minimaxity. We note significant differences between original quantile curve and the optimal projected one for veterans. For example, the median of the annual earnings for veterans is 9,819 dollars according to the original estimate and 9,929 dollars according to the projected estimate. The maximal difference of 1,767 dollars occurs at the 0.725 quantile. In contrast, we find less difference between the original structural quantile function and the optimal projected counterpart for the non-veterans, with the maximal gap being 403 dollars at the 0.725 quantile.

5 Conclusion

In this paper, we have derived the local asymptotic minimax lower bound for a class of plug-in estimators of directionally differentiable parameters, which arise in a large class of econometric problems. The employment of minimaxity criterion, although perhaps not fully necessary, seems to the most suitable one for our purposes. The derived lower bound is intrinsically complicated. Nonetheless, we have been able to present a general construction procedure to show attainability of the lower bound.

Appendix A Proofs of Main Results

The following list includes notation and definitions that will be used in the appendix.

A^ϵ	For a set A in a metric space (T,d) and $\epsilon>0$, $A^\epsilon \equiv \{a \in T : d(a,A) \le \epsilon\}$.
$A \subset_f B$	For sets A and B, A is a finite subset of B.
$M^\intercal$	For an $m \times n$ matrix M, $M^\intercal$ is the transpose of M.
K^m_λ	For $\lambda>0$, $K^m_\lambda \equiv \{x \in \mathbf{R}^m : \|x\| \le \lambda\}$.
$[A]$	For a set A in a normed space, $[A]$ is the closed linear span of A.
$A^\perp$	For a set A in a Hilbert space $\mathbb{H}$, $A^\perp \equiv \{x \in \mathbb{H} : \langle x,y\rangle_{\mathbb{H}} = 0 \ , \forall\, y \in A\}$.
$\ell^\infty(T)$	The space of bounded functions on T.
$\|f\|_{L^p}$	For a measure space $(T,\mathcal{M},\mu)$ and $1 \le p < \infty$, $\|f\|_{L^p} \equiv \{\int |f|^p \, d\mu\}^{1/p}$.
$L^p(T)$	For a measure space $(T,\mathcal{M},\mu)$, $L^p(T) \equiv \{f : T \to \mathbf{R} : \|f\|_{L^p} < \infty\}$.
$\mathrm{BL}_1(T)$	The set of functions f with $\sup_{t\in T}\|f(t)\| \le 1$ and $\|f(t_1)-f(t_2)\| \le d(t_1,t_2)$.

Proof of Theorem 3.1: For each finite subset $I \subset H$, we have

$$\liminf_{n\to\infty} \sup_{h\in I} E_{n,h}[\ell(r_n\{\phi(\hat\theta_n) - \phi(\theta_n(h))\})] \\ \ge \sup_{h\in I} \liminf_{n\to\infty} E_{n,h}[\ell(r_n\{\phi(\hat\theta_n) - \phi(\theta_n(h))\})] \ . \tag{A.1}$$

By Assumption 3.3, ℓ is continuous and positive. In turn, Lemma A.1 allows us to invoke the portmanteau theorem to conclude that

$$\liminf_{n\to\infty} E_{n,h}[\ell(r_n\{\phi(\hat\theta_n) - \phi(\theta_n(h))\})] \\ \ge E[\ell(\phi'_{\theta_0}(\mathbb{G} + \theta'_0(h) + \Delta) - \phi'_{\theta_0}(\theta'_0(h) + \Delta))] \ . \tag{A.2}$$

Combining results (A.1) and (A.2) we thus have

$$\liminf_{n\to\infty} \sup_{h\in I} E_{n,h}[\ell(r_n\{\phi(\hat\theta_n) - \phi(\theta_n(h))\})] \\ \ge \sup_{h\in I} E[\ell(\phi'_{\theta_0}(\mathbb{G} + \theta'_0(h) + \Delta) - \phi'_{\theta_0}(\theta'_0(h) + \Delta))] \ . \tag{A.3}$$

Taking supremum on both sides in (A.3) over all finite $I \subset H$ yields that

$$\begin{aligned} \sup_{I\subset_f H} \liminf_{n\to\infty} \sup_{h\in I} & E_{n,h}[\ell(r_n\{\phi(\hat\theta_n) - \phi(\theta_n(h))\})] \\ &\ge \sup_{I\subset_f H} \sup_{h\in I} E[\ell(\phi'_{\theta_0}(\mathbb{G} + \theta'_0(h) + \Delta) - \phi'_{\theta_0}(\theta'_0(h) + \Delta))] \\ &= \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G} + \theta'_0(h) + \Delta) - \phi'_{\theta_0}(\theta'_0(h) + \Delta))] \\ &= \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G} + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \ , \end{aligned} \tag{A.4}$$

where the last equality is due to the fact that $\Delta \in \theta'_0(H)$ by Assumption 3.1 and the fact that H is linear by Assumption 2.1(i).

In view of (A.4) and the desired lower bound in (3.10), it suffices to show that,

$$\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G} + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \\ \ge \inf_{u\in\mathbb{D}} \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \ . \tag{A.5}$$

Towards this end, we follow the idea of Song (2014a) but, instead of employing the purification theorem initially developed by Dvoretzky et al. (1950, 1951), we appeal to a more generalized version in Feinberg and Piunovskiy (2006) and hence are able to simplify the proof that would be otherwise involved.

Since H is separable by Assumption 2.1(i), we may pick a sequence $\{h_j\}_{j=1}^{\infty}$ that is dense in H. By positivity and continuity of ℓ implied by Assumption 3.3 and continuity of θ_0' and ϕ_{θ_0}' implied by Assumptions 2.2 and 3.4, we may conclude by Fatou's lemma that $E[\ell(\phi_{\theta_0}'(\mathbb{G}+\theta_0'(h))-\phi_{\theta_0}'(\theta_0'(h)))]$ is lower semicontinuous in h. It follows by Lemma A.5 that

$$\sup_{h\in H} E[\ell(\phi_{\theta_0}'(\mathbb{G}+\theta_0'(h))-\phi_{\theta_0}'(\theta_0'(h)))] = \sup_{j\in\mathbf{N}} E[\ell(\phi_{\theta_0}'(\mathbb{G}+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))] \ . \tag{A.6}$$

Fix $J \in \mathbf{N}$. For $j = 1, \ldots, J$, write $\rho(z,u) = (\rho_1(z,u), \ldots, \rho_J(z,u))^\intercal$ where

$$\rho_j(z,u) = E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))]z \ .$$

By Assumptions 2.1, 2.2 and 2.3, Theorem 2.1 applies so that we may write $\mathbb{G} \stackrel{d}{=} \mathbb{G}_0+\mathbb{U}$, where $\mathbb{G}_0$ is the efficient Gaussian component and $\mathbb{U}$ is the noise term independent of $\mathbb{G}_0$. Denote the distribution of $\mathbb{U}$ by Q. For fixed $\lambda > 1$, let Z follow the uniform distribution ν_λ supported on $[1,\lambda]$. By Theorem 1 in Feinberg and Piunovskiy (2006), there is a measurable map $u^* : [1,\lambda] \to \mathbb{D}$ such that

$$\int_1^\lambda \int_{\mathbb{D}} \rho(z,u)\, Q(du)\nu_\lambda(dz) = \int_1^\lambda \rho(z,u^*(z))\, \nu_\lambda(dz) \ ,$$

which in turn implies that, for all $j = 1, \ldots, J$,

$$\begin{aligned}
&\frac{1+\lambda}{2}\int_{\mathbb{D}} E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))]\, Q(du) \\
&\qquad = \int_1^\lambda E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u^*(z)+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))]z\, \nu_\lambda(dz) \\
&\qquad \geq \int_1^\lambda E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u^*(z)+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))]\, \nu_\lambda(dz) \ ,
\end{aligned} \tag{A.7}$$

where the inequality exploits the facts that $z \geq 1$ and that $\ell \geq 0$. By change of variable applied to the right hand side of (A.7), we have for all $j = 1, \ldots, J$,

$$\begin{aligned}
&\frac{1+\lambda}{2}\int_{\mathbb{D}} E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))]\, Q(du) \\
&\qquad \geq \int_0^1 E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u^*((\lambda-1)y+1)+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))]\, dy \ .
\end{aligned} \tag{A.8}$$

It follows that

$$\begin{aligned}
&\frac{1+\lambda}{2}\max_{j=1,\ldots,J} E[\ell(\phi_{\theta_0}'(\mathbb{G}+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))] \\
&\qquad \geq \inf_{\lambda>1}\max_{j=1,\ldots,J}\int_0^1 E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u^*((\lambda-1)y+1)+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))]\, dy \\
&\qquad \geq \inf_{u\in\mathcal{R}(u^*)}\max_{j=1,\ldots,J}\int_0^1 E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))]\, dy \\
&\qquad \geq \inf_{u\in\mathbb{D}}\max_{j=1,\ldots,J} E[\ell(\phi_{\theta_0}'(\mathbb{G}_0+u+\theta_0'(h_j))-\phi_{\theta_0}'(\theta_0'(h_j)))] \ ,
\end{aligned} \tag{A.9}$$

where $\mathcal{R}(u^*)$ denotes the range of u^*.

Letting $\lambda \downarrow 1$ and then $J \to \infty$ in (A.9) yields

$$\sup_{j\in\mathbf{N}} E[\ell(\phi'_{\theta_0}(\mathbb{G} + \theta'_0(h_j)) - \phi'_{\theta_0}(\theta'_0(h_j)))] \geq \inf_{u\in\mathbb{D}} \sup_{j\in\mathbf{N}} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h_j)) - \phi'_{\theta_0}(\theta'_0(h_j)))] \ . \tag{A.10}$$

Combining (A.6), (A.10), and the fact that the expectation on the right hand side is also lower semicontinuous in h by Assumptions 2.2, 3.2 and 3.3, we thus conclude that

$$\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G} + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \geq \inf_{u\in\mathbb{D}} \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \ , \tag{A.11}$$

proving (A.5) and hence the Theorem. ■

Lemma A.1 *Let $(\mathscr{X}_n, \mathcal{A}_n, \{P_{n,h} : h \in H\})$ be a sequence of statistical experiments, and $\hat{\theta}_n$ be an estimator for the parameter $\theta : \{P_{n,h}\} \to \mathbb{D}$. If Assumptions 2.2, 2.3, 3.1 and 3.2 hold, then*

$$r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(h))\} \overset{L_{n,h}}{\to} \phi'_{\theta_0}(\mathbb{G} + \theta'_0(h) + \Delta) - \phi'_{\theta_0}(\theta'_0(h) + \Delta) \tag{A.12}$$

for every $h \in H$.

Proof: Rewrite

$$r_n\{\phi(\hat{\theta}_n) - \phi(\theta_n(h))\} = r_n\{\phi(\hat{\theta}_n) - \phi(\theta_0)\} - r_n\{\phi(\theta_n(h)) - \phi(\theta_0)\} \ . \tag{A.13}$$

By Assumptions 2.3, 2.2, and 3.1, we have

$$r_n(\hat{\theta}_n - \theta_0) = r_n\{\hat{\theta}_n - \theta_n(h)\} + r_n\{\theta_n(h) - \theta_n(0)\} + r_n\{\theta_n(0) - \theta_0\} \overset{L_{n,h}}{\to} \mathbb{G} + \theta'_0(h) + \Delta \ ,$$

for every $h \in H$. By Assumption 3.2, ϕ is Hadamard directionally differentiable at θ_0 tangentially to $\mathbb{D}$, and hence by the Delta method (Fang and Santos, 2014, Theorem 2.1) we may conclude that

$$r_n\{\phi(\hat{\theta}_n) - \phi(\theta_0)\} \overset{L_{n,h}}{\to} \phi'_{\theta_0}(\mathbb{G} + \theta'_0(h) + \Delta) \ . \tag{A.14}$$

On the other hand, Assumptions 2.2, and 3.1 imply that for all $h \in H$,

$$r_n\{\theta_n(h) - \theta_0\} = r_n\{\theta_n(h) - \theta_n(0)\} + r_n\{\theta_n(0) - \theta_0\} \to \theta'_0(h) + \Delta \ ,$$

whence by Assumption 3.2,

$$r_n\{\phi(\theta_n(h)) - \phi(\theta_0)\} \to \phi'_{\theta_0}(\theta'_0(h) + \Delta) \ . \tag{A.15}$$

The Lemma then follows from displays (A.13), (A.14) and (A.15). ■

Proof of Lemma 3.1: By Assumption 3.4, we have

$$\|\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h))\|_{\mathbb{E}} \leq C_{\phi'} \|\mathbb{G}_0 + u\|_{\mathbb{D}} \ ,$$

and hence by ρ being nondecreasing

$$\begin{aligned}\inf_{u\in\mathbb{D}}\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))] \\ &= \inf_{u\in\mathbb{D}}\sup_{h\in H} E[\rho(\|\phi'_{\theta_0}(\mathbb{G}_0+u+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h))\|_{\mathbb{E}})] \\ &\leq \inf_{u\in\mathbb{D}} E[\rho(C_{\phi'}\|\mathbb{G}_0+u\|_{\mathbb{D}})] \ . \end{aligned} \tag{A.16}$$

For each $c \geq 0$, the set $A_c \equiv \{y : \rho(C_{\phi'}\|y\|_{\mathbb{D}}) \leq c\}$ is clearly symmetric. It is also closed since if $\{y_n\} \subset A_c$ and $y_n \to y$, then ρ being lower semicontinuous implies that

$$\rho(C_{\phi'}\|y\|_{\mathbb{D}}) \leq \liminf_{n\to\infty} \rho(C_{\phi'}\|y_n\|_{\mathbb{D}}) \leq c \ .$$

Finally, A_c is convex since if $y_1, y_2 \in A_c$, then for any $\lambda \in (0,1)$

$$\begin{aligned}\rho(C_{\phi'}\|\lambda y_1+(1-\lambda)y_2\|_{\mathbb{D}}) \leq &\rho(\lambda C_{\phi'}\|y_1\|_{\mathbb{D}}+(1-\lambda)C_{\phi'}\|y_2\|_{\mathbb{D}}) \\ &\leq \rho(\max\{C_{\phi'}\|y_1\|_{\mathbb{D}}, C_{\phi'}\|y_2\|_{\mathbb{D}}\}) \leq c \ .\end{aligned}$$

Therefore $\rho(C_{\phi'}\|\cdot\|_{\mathbb{D}})$ is subconvex. We thus conclude from result (A.16) and Anderson's lemma (van der Vaart and Wellner, 1996) that

$$\begin{aligned}\inf_{u\in\mathbb{D}}\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))] \\ &\leq \inf_{u\in\mathbb{D}} E[\rho(C_{\phi'}\|\mathbb{G}_0+u\|_{\mathbb{D}})] = E[\rho(C_{\phi'}\|\mathbb{G}_0\|_{\mathbb{D}})] < \infty \ .\end{aligned}$$

This establishes the Lemma. ■

Proof of Corollary 3.1: By Theorem 3.1, we know that the lower bound is given by

$$\inf_{u\in\mathbb{D}}\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))] \ . \tag{A.17}$$

By Assumptions 2.2, 3.4 and 3.3, we may conclude by Fatou's lemma that the expectation in (A.17) is lower semicontinuous in h. It follows by Lemma A.5 that

$$\inf_{u\in\mathbb{D}}\sup_{c\in\overline{\theta'_0(H)}} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u+c)-\phi'_{\theta_0}(c))] \ . \tag{A.18}$$

Since $\mathbb{G}_0$ is Gaussian in $\mathbb{D} \equiv \mathbf{R}^m$ with nonsingular covariance Σ_0, by Theorem 2.1 it must be the case that $\overline{\theta'_0(H)} = \mathbf{R}^m$. The Corollary then follows. ■

Proof of Theorem 3.2: Suppose first that the loss function ℓ is bounded by $M > 0$. Fix $\epsilon > 0$. Then there is some $u_\epsilon \in \mathbb{D}$ such that

$$\begin{aligned}\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u_\epsilon+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))] \\ &\leq \inf_{u\in\mathbb{D}}\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))]+\frac{\epsilon}{4} \ .\end{aligned} \tag{A.19}$$

By Assumptions 3.3 and 3.4, $\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))]$ is (Lipschitz) continuous in u. Thus, there is some $\delta > 0$ such that

$$\begin{aligned}\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))] \\ &\leq \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+u_\epsilon+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))]+\frac{\epsilon}{4} \ ,\end{aligned}$$

whenever $\|u - u_\epsilon\|_{\mathbb{D}} < \delta$. By Assumption 3.6(iii) and the fact that $\tau_k \to \infty$ as $k \to \infty$, there is some $v_k \in K^k_{\tau_k}$ such that $\|u_k - u_\epsilon\|_{\mathbb{D}} < \delta$ with $u_k \equiv (\psi^k)^\intercal v_k$ for all k large enough, which in turn means that

$$\begin{aligned}&\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u_k + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \\ &\qquad \le \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u_\epsilon + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] + \frac{\epsilon}{4} \ . \end{aligned} \tag{A.20}$$

Combining results (A.19) and (A.20) we thus have for all k large enough,

$$\begin{aligned}&\inf_{v\in K^k_{\tau_k}} \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \\ &\qquad \le \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u_k + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \\ &\qquad \le \inf_{u\in\mathbb{D}} \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] + \frac{\epsilon}{2} \ . \end{aligned} \tag{A.21}$$

Next, for notational simplicity, define

$$B_m(v) \equiv \max_{c\in K^m_{\lambda_m}} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(g^m)^\intercal c) - \phi'_{\theta_0}(\theta'_0(g^m)^\intercal c))] \ ,$$

$$B(v) \equiv \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \ , \ \Psi_{k,m} \equiv \operatorname*{arg\,min}_{v\in K^k_{\tau_k}} B_m(v) \ .$$

Fix k large enough so that (A.21) holds. By Assumptions 3.3 and 3.4, it is clear that both $B(v)$ and $B_m(v)$ for each $m \in \mathbf{N}$ are continuous functions on $K^k_{\tau_k}$. Moreover, $B_m(v)$ increasingly converges to $B(v)$ as $m \to \infty$ for each $v \in K^k_{\tau_k}$ with additional Assumption 3.6(i). To see this, fix $v \in K^k_{\tau_k}$ and pick $h_\epsilon \in H$ such that

$$B(v) - \epsilon \le E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(h_\epsilon)) - \phi'_{\theta_0}(\theta'_0(h_\epsilon)))] \le B(v) \ . \tag{A.22}$$

By Assumptions 2.2, 3.3 and 3.4, $E[\ell(\phi'_{\theta_0}(\mathbb{G}_0+(\psi^k)^\intercal v+\theta'_0(h))-\phi'_{\theta_0}(\theta'_0(h)))]$ is (Lipschitz) continuous in h, and hence by Assumption 3.6(i) and the fact that $\lambda_m \to \infty$ as $m \to \infty$, we have for all m sufficiently large

$$E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(h_\epsilon)) - \phi'_{\theta_0}(\theta'_0(h_\epsilon)))] \le B_m(v) + \epsilon \ . \tag{A.23}$$

Combining previous two displays we obtain $B(v) - 2\epsilon \le B_m(v) \le B(v)$ for all m sufficiently large. This shows that $B_m(v)$ increasingly converges to $B(v)$ for each $v \in K^k_{\tau_k}$. It follows by Dini's theorem (Aliprantis and Border, 2006, Theorem 2.66) that $B_m \to B$ uniformly on $K^k_{\tau_k}$. We thus conclude that there is an m_0 such that for all $m \ge m_0$, $B(v) \le B_m(v) + \epsilon/2$ or equivalently

$$\begin{aligned}&\sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \\ &\qquad \le \sup_{c\in K^m_{\lambda_m}} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(g^m)^\intercal c) - \phi'_{\theta_0}(\theta'_0(g^m)^\intercal c))] + \frac{\epsilon}{2} \ , \end{aligned} \tag{A.24}$$

for all $v \in K^k_{\tau_k}$.

Next, fix an arbitrary subsequence $\{n_\ell\}$. For fixed $m, k \in \mathbf{N}$, $B_m(\cdot)$ is continuous on $K^k_{\tau_k}$ by Assumptions 3.3 and 3.4, which, together with compactness of $K^k_{\tau_k}$, implies

that $\Psi_{k,m}$ is nonempty and compact by Theorem 2.43 in Aliprantis and Border (2006). Combination of Lemma A.2 and Lemma A.6 then implies that there exist a further subsequence $\{n_{\ell_j}\}$ and some $v^*_{k,m} \in \Psi_{k,m}$ such that

$$\hat{v}_{n_{\ell_j},k,m} \xrightarrow{p} v^*_{k,m} \ , \tag{A.25}$$

as $j \to \infty$ under $\{P_{n,h}\}$ with $h \in H$, for each $k, m \in \mathbf{N}$. Result (A.25), together with Assumptions 3.7, 2.2, 3.1 and 3.2, allows us to invoke Slutsky's theorem and the Delta method to conclude that

$$r_{n_{\ell_j}}\{\phi(\hat{\theta}^*_{n_{\ell_j}} + \frac{\hat{u}_{n_{\ell_j},k,m}}{r_{n_{\ell_j}}}) - \phi(\theta_{n_{\ell_j}}(h))\} \overset{L_{n_{\ell_j},h}}{\to} \phi'_{\theta_0}(\mathbb{G}_0 + u^*_{k,m} + \Delta + \theta'_0(h)) - \phi'_{\theta_0}(\Delta + \theta'_0(h))$$

for each $h \in H$, where $u^*_{k,m} \equiv (\psi^k)^\intercal v^*_{k,m}$. Since ℓ is bounded and continuous, it follows that for all m sufficiently large and all $k \in \mathbf{N}$,

$$\begin{aligned}
&\sup_{I \subset_f H} \limsup_{j\to\infty} \sup_{h\in I} E_{n_{\ell_j},h}[\ell(r_{n_{\ell_j}}\{\phi(\hat{\theta}_{n_{\ell_j}} + \frac{\hat{u}_{n_{\ell_j},k,m}}{r_{n_{\ell_j}}}) - \phi(\theta_{n_{\ell_j}}(h))\})] \\
&\qquad = \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v^*_{k,m} + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] \\
&\qquad \le \max_{v\in\Psi_{k,m}} B(v) \le \max_{v\in\Psi_{k,m}} B_m(v) + \frac{\epsilon}{2} = \inf_{v\in K^k_{\tau_k}} B_m(v) + \frac{\epsilon}{2} \ ,
\end{aligned}$$

where the first inequality is due to $v^*_{k,m} \in \Psi_{k,m}$, the second inequality is by result (A.24), while the last equality is by definition of $\Psi_{k,m}$. This implies that for all k large enough,

$$\begin{aligned}
&\limsup_{m\to\infty} \sup_{I\subset_f H} \limsup_{j\to\infty} \sup_{h\in I} E_{n_{\ell_j},h}[\ell(r_{n_{\ell_j}}\{\phi(\hat{\theta}_{n_{\ell_j}} + \frac{\hat{v}_{n_{\ell_j},k,m}}{r_{n_{\ell_j}}}) - \phi(\theta_{n_{\ell_j}}(h))\})] \\
&\qquad \le \limsup_{m\to\infty} \inf_{v\in K^k_{\tau_k}} B_m(v) + \frac{\epsilon}{2} = \inf_{v\in K^k_{\tau_k}} B(v) + \frac{\epsilon}{2} \\
&\qquad \le \inf_{u\in\mathbb{D}} \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] + \epsilon \ ,
\end{aligned} \tag{A.26}$$

where the equality follows from the fact that $B_m \to B$ uniformly on $K^k_{\tau_k}$, and the last inequality holds for all k sufficiently large due to (A.21). We thus have

$$\begin{aligned}
&\limsup_{k\to\infty} \limsup_{m\to\infty} \sup_{I\subset_f H} \limsup_{j\to\infty} \sup_{h\in I} E_{n_{\ell_j},h}[\ell(r_{n_{\ell_j}}\{\phi(\hat{\theta}_{n_{\ell_j}} + \frac{\hat{v}_{n_{\ell_j},k,m}}{r_{n_{\ell_j}}}) - \phi(\theta_{n_{\ell_j}}(h))\})] \\
&\qquad \le \inf_{u\in\mathbb{D}} \sup_{h\in H} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + \theta'_0(h)) - \phi'_{\theta_0}(\theta'_0(h)))] + \epsilon \ .
\end{aligned}$$

The theorem then follows for bounded ℓ by the facts that $\{n_\ell\}$ and ϵ are arbitrary. For general loss functions ℓ, replace ℓ in the above proof with $\ell_M \equiv \ell \wedge M$ and then let $M \to \infty$. ■

Lemma A.2 *Suppose that Assumptions 2.1, 2.2, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6 and 3.7 hold. Let $\hat{v}_{n,k,m} \in \hat{\Psi}_{k,m}$. If the loss function ℓ is bounded, then it follows that for each $k, m \in \mathbf{N}$,*

$$d(\hat{v}_{n,k,m}, \Psi_{k,m}) \xrightarrow{p} 0 \ , \tag{A.27}$$

under $P_{n,h}$ for all $h \in H$.

PROOF: Fix $k, m \in \mathbf{N}$ throughout. For notational simplicity, write $\vartheta \equiv (v^\intercal, c^\intercal)^\intercal \in \Theta \equiv K^k_{\tau_k} \times K^m_{\lambda_m}$ and $\eta_0 \equiv (\theta'_0, \phi'_{\theta_0})$, and define the function $f_{\vartheta,\eta_0}(\cdot) : \mathbb{D} \to \mathbf{R}$ by

$$f_{\vartheta,\eta_0}(z) \equiv \ell(\phi'_{\theta_0}(z + (\psi^k)^\intercal v + \theta'_0(g^m)^\intercal c) - \phi'_{\theta_0}(\theta'_0(g^m)^\intercal c)) \ .$$

Let $\hat{\eta}_n \equiv (\hat{\theta}'_n, \hat{\phi}'_n)$ and define

$$\begin{aligned}
Pf_{\vartheta,\eta_0} &\equiv E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(g^m)^\intercal c) - \phi'_{\theta_0}(\theta'_0(g^m)^\intercal c))] \ , \\
Pf_{\vartheta,(\theta'_0,\hat{\phi}'_n)} &\equiv E[\ell(\hat{\phi}'_n(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta'_0(g^m)^\intercal c) - \hat{\phi}'_n(\theta'_0(g^m)^\intercal c))|\{X_i\}] \ , \\
Pf_{\vartheta,\hat{\eta}_n} &\equiv E[\ell(\hat{\phi}'_n(\mathbb{G}_0 + (\psi^k)^\intercal v + \hat{\theta}'_n(\hat{g}^m)^\intercal c) - \hat{\phi}'_n(\hat{\theta}'_n(\hat{g}^m)^\intercal c))|\{X_i\}] \ , \\
\mathbb{P}_n f_{\vartheta,\hat{\eta}_n} &\equiv E[\ell(\hat{\phi}'_n(\hat{\mathbb{G}}^*_n + (\psi^k)^\intercal v + \hat{\theta}'_n(\hat{g}^m)^\intercal c) - \hat{\phi}'_n(\hat{\theta}'_n(\hat{g}^m)^\intercal c))|\{X_i\}] \ ,
\end{aligned}$$

where $Pf_{\vartheta,(\theta'_0,\hat{\phi}'_n)}$, and $Pf_{\vartheta,\hat{\eta}_n}$ are expectations taken with respect to $\mathbb{G}_0$ while holding $\{X_i\}_{i=1}^n$ fixed. The ensuing arguments are organized parallel to those of the consistency result in the theory of the extremum estimation, the only difference being that the set of population minimizers is possibly a nonsingleton. Therefore, we need to show a uniform convergence result and an identification condition.

Uniform Convergence: For each $\epsilon > 0$,

$$\sup_{\vartheta \in \Theta} |\mathbb{P}_n f_{\vartheta,\hat{\eta}_n} - Pf_{\vartheta,\eta_0}| = o_p(1) \ , \tag{A.28}$$

under $\{P_{n,0}\}$. In turn, it suffices to show that

$$\sup_{\vartheta \in \Theta} |\mathbb{P}_n f_{\vartheta,\hat{\eta}_n} - Pf_{\vartheta,\hat{\eta}_n}| = o_p(1) \ , \tag{A.29a}$$

$$\sup_{\vartheta \in \Theta} |Pf_{\vartheta,\hat{\eta}_n} - Pf_{\vartheta,(\theta'_0,\hat{\phi}'_n)}| = o_p(1) \ , \tag{A.29b}$$

$$\sup_{\vartheta \in \Theta} |Pf_{\vartheta,(\theta'_0,\hat{\phi}'_n)} - Pf_{\vartheta,\eta_0}| = o_p(1) \ , \tag{A.29c}$$

under $\{P_{n,0}\}$. Fix $\epsilon > 0$ and consider (A.29a). Note that for every realization of $\{X_i\}$, the real valued functions

$$\ell(\hat{\phi}'_n(\cdot + (\psi^k)^\intercal v + \hat{\theta}'_n(\hat{g}^m)^\intercal c) - \hat{\phi}'_n(\hat{\theta}'_n(\hat{g}^m)^\intercal c))$$

are bounded and Lipschitz continuous on $\mathbb{D}$ with Lipschitz constant $C_\ell C_{\hat{\phi}'}$ by Assumptions 3.3 and 3.5(iii)-(b). It then follows by Assumption 3.5(i) that

$$\sup_{\vartheta \in \Theta} |\mathbb{P}_n f_{\vartheta,\hat{\eta}_n} - Pf_{\vartheta,\hat{\eta}_n}| \leq \sup_{f \in \mathrm{BL}_a(\mathbb{D})} |E[f(\hat{\mathbb{G}}^*_n)|\{X_i\}] - E[f(\mathbb{G}_0)]| = o_p(1) \ ,$$

under $\{P_{n,0}\}$, where $a \equiv \max\{M, C_\ell C_{\hat{\phi}'}\}$ with M being a upper bound of ℓ, proving (A.29a).

Next, consider (A.29b). We have

$$\begin{aligned}
\sup_{\vartheta \in \Theta} |Pf_{\vartheta,\hat{\eta}_n} - Pf_{\vartheta,(\theta'_0,\hat{\phi}'_n)}| &\leq 2C_\ell C_{\hat{\phi}'} \|\hat{\theta}'_n(\hat{g}^m)^\intercal c - \theta'_0(g^m)^\intercal c\|_{\mathbb{D}} \\
&\leq 2C_\ell C_{\hat{\phi}'} \lambda_m \sum_{j=1}^m \|\hat{\theta}'_n(\hat{g}_j) - \theta'_0(g_j)\|_{\mathbb{D}} = o_p(1) \ ,
\end{aligned}$$

under $\{P_{n,0}\}$, where the first inequality is due to Assumptions 3.3 and 3.5(iii)-(b), and the second inequality is by Assumptions 3.5(ii) and 3.6(ii). This shows (A.29b).

Lastly, let us deal with (A.29c). For fixed $k, m \in \mathbf{N}$, $K_1 \equiv \{(\psi^k)^\intercal v : v \in K^k_{\tau_k}\}$ and $K_2 \equiv \{\theta_0'(g^m)^\intercal c : c \in K^m_{\lambda_m}\}$ is compact in $\mathbb{D}$ by Proposition 4.26 in Folland (1999). Fix $\epsilon, \eta > 0$. Since $\mathbb{G}_0$ is tight, there is some compact $K_0 \subset \mathbb{D}$ such that $P(\mathbb{G}_0 \notin K_0) < \eta/(2M)$. Let $K \equiv K_0 + K_1 + K_2$. Clearly, $K \subset \mathbb{D}$ is compact. We now have

$$\begin{aligned}
\sup_{\vartheta\in\Theta} |Pf_{\vartheta,(\theta_0',\hat{\phi}_n')} - Pf_{\vartheta,\eta_0}| \leq \sup_{\vartheta\in\Theta} E[|\ell(\hat{\phi}_n'(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta_0'(g^m)^\intercal c) \\
- \hat{\phi}_n'(\theta_0'(g^m)^\intercal c)) - \ell(\phi_{\theta_0}'(\mathbb{G}_0 + (\psi^k)^\intercal v + \theta_0'(g^m)^\intercal c) - \phi_{\theta_0}'(\theta_0'(g^m)^\intercal c))|\,|\{X_i\}] \\
\leq \sup_{z\in K} C_\ell \|\hat{\phi}_n'(z) - \phi_{\theta_0}'(z)\|_{\mathbb{E}} + \sup_{z\in K_2} C_\ell \|\hat{\phi}_n'(z) - \phi_{\theta_0}'(z)\|_{\mathbb{E}} + 2M \cdot P(\mathbb{G}_0 \notin K_0) \\
\leq o_p(1) + \eta \ ,
\end{aligned}$$

where the first inequality is by the triangle inequality, the second is by Assumption 3.3, and the last is by Lemma A.7. This immediately implies (A.29c) and hence we conclude that (A.28) holds.

Identification Condition: For each $\epsilon > 0$,

$$\inf_{v\in K^k_{\tau_k}\setminus\Psi^\epsilon_{k,m}} \sup_{c\in K^m_{\lambda_m}} Pf_{(v,c),\eta_0} > \inf_{v\in K^k_{\tau_k}} \sup_{c\in K^m_{\lambda_m}} Pf_{(v,c),\eta_0} \ , \tag{A.30}$$

or equivalently, $\inf_{v\in K^k_{\tau_k}\setminus\Psi^\epsilon_{k,m}} B_m(v) > \inf_{v\in K^k_{\tau_k}} B_m(v)$, where $\Psi^\epsilon_{k,m} \equiv \{v \in \mathbf{R}^k : d(v, \Psi_{k,m}) \leq \epsilon\}$. To see this, fix $\epsilon > 0$ and suppose that $\inf_{u\in K^k_{\tau_k}\setminus\Psi^\epsilon_{k,m}} B_m(v) = \inf_{u\in K^k_{\tau_k}} B_m(v)$. Then we may pick a sequence $\{v_i\} \subset K^k_{\tau_k}\setminus\Psi^\epsilon_{k,m}$ such that

$$B_m(v_i) \to \inf_{v\in K^k_{\tau_k}} B_m(v) \ \text{ as } \ i \to \infty \ .$$

By passing to a subsequence if necessary, we may assume that $v_i \to v^*$ as $i \to \infty$ where $v^* \in \overline{K^k_{\tau_k}\setminus\Psi^\epsilon_{k,m}}$. Assumptions 3.3 and 3.4 imply that $B_m(v)$ is (Lipschitz) continuous in v and therefore

$$B_m(v^*) = \inf_{v\in K^k_{\tau_k}} B_m(v) \ ,$$

meaning that $v^* \in \Psi_{k,m}$. On the other hand, $v^* \in \overline{K^k_{\tau_k}\setminus\Psi^\epsilon_{k,m}}$ implies that we may take a sequence $\{v_j\} \subset K^k_{\tau_k}\setminus\Psi^\epsilon_{k,m}$ such that $v_j \to v^*$ as $j \to \infty$, which in turn implies that

$$d(v^*, \Psi_{k,m}) = \lim_{j\to\infty} d(v_j, \Psi_{k,m}) \geq \epsilon > 0 \ ,$$

a contradiction. Therefore, (A.30) must hold.

We are now in a position to show result (A.27). Fix $\epsilon > 0$. By the identification result (A.30), there is some $\delta > 0$ such that whenever $v \in K^k_{\tau_k}\setminus\Psi^\epsilon_{k,m}$ we have

$$B_m(v) - B_m(v^*_{k,m}) \geq \delta \ , \tag{A.31}$$

where $v^*_{k,m}$ is any element in $\Psi_{k,m}$. It follows that

$$\begin{aligned}
P_{n,0}(d(\hat{v}_{n,k,m},\Psi_{k,m}) > \epsilon) &= P_{n,0}(\hat{v}_{n,k,m} \in K^k_{\tau_k} \backslash \Psi^\epsilon_{k,m}) \\
&\leq P_{n,0}(B_m(\hat{v}_{n,k,m}) - B_m(v^*_{k,m}) \geq \delta) \\
&= P_{n,0}(B_m(\hat{v}_{n,k,m}) - \hat{B}_m(\hat{v}_{n,k,m}) + \hat{B}_m(\hat{v}_{n,k,m}) - B_m(v^*_{k,m}) \geq \delta) \\
&\leq P_{n,0}(B_m(\hat{v}_{n,k,m}) - \hat{B}_m(\hat{v}_{n,k,m}) + \hat{B}_m(v^*_{k,m}) + \epsilon_n - B_m(v^*_{k,m}) \geq \delta) \\
&\leq P_{n,0}(2 \sup_{v \in K^k_{\tau_k}} |B_m(v) - \hat{B}_m(v)| + \epsilon_n \geq \delta) \\
&\leq P_{n,0}(2 \sup_{\vartheta \in \Theta} |\mathbb{P}_n f_{\vartheta,\hat{\eta}_n} - P f_{\vartheta,\eta_0}| + \epsilon_n \geq \delta) \to 0 \ , \qquad \text{(A.32)}
\end{aligned}$$

where the second inequality is by the definition of $\hat{v}_{n,k,m}$, and the last step is by (A.28) and the fact that $\epsilon_n = o_p(1)$ as $n \to \infty$. By Assumption 2.1(ii), $P_{n,h}$ and $P_{n,0}$ are mutually contiguous for each $h \in H$; see, for example, Example 6.5 in van der Vaart (1998). Result (A.27) then follows from (A.32) and Le Cam's first lemma. ■

Proof of Theorem 3.3: The proof follows closely that of Theorem 3.2. Define

$$B(u) \equiv \max_{c \in \mathbf{R}^m} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c))] \ ,$$
$$B_\lambda(u) \equiv \max_{c \in K^m_\lambda} E[\ell(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c))] \ , \ \Psi_{\tau,\lambda} \equiv \arg\min_{u \in K^m_\tau} B_\lambda(u) \ .$$

Again, consider first the case when ℓ is bounded. Fix $\epsilon > 0$ and $\tau > 0$. By Assumption 3.3 and 3.4, $B_\lambda(u)$ and $B(u)$ are both (Lipschitz) continuous in u. Moreover, it is clear that $B_\lambda(u) \uparrow B(u)$ as $\lambda \uparrow \infty$ for each $u \in K^m_\tau$. It then follows by Dini's theorem that $B_\lambda \to B$ uniformly on K^m_τ so that we may find some $\lambda > 0$ with $\lambda \geq \tau$ if necessary such that

$$B(u) \leq B_\lambda(u) + \epsilon \text{ for all } u \in K^m_\tau \ . \qquad \text{(A.33)}$$

The rest of the proof is essentially the same as that of Theorem 3.3 by employing subsequence arguments, in view of Lemma A.3 and Lemma A.6. ■

Lemma A.3 *Suppose that Assumptions 2.1, 2.2, 2.3 3.1, 3.2, 3.3, 3.4, and 3.5(i)(iii) hold. Let $\hat{u}_{n,\tau,\lambda} \in \hat{\Psi}_{\tau,\lambda}$. Further assume that the loss function ℓ is bounded. Then for all $\tau, \lambda > 0$ we have*

$$d(\hat{u}_{n,\tau,\lambda}, \Psi_{\tau,\lambda}) \xrightarrow{p} 0 \ , \qquad \text{(A.34)}$$

under $P_{n,h}$ for each $h \in H$.

Proof: Following the proof of Lemma A.2, it suffices to show a unform convergence condition and an identification condition. Since the identification condition can be shown using exactly the same arguments as before, we shall only prove the following: for fixed $\tau, \lambda > 0$,

$$\sup_{u \in K^m_\tau} |\hat{B}_\lambda(u) - B_\lambda(u)| = o_p(1) \ , \qquad \text{(A.35)}$$

under $P_{n,0}$. To this end, rewrite

$$\begin{aligned}
&\sup_{u\in K_\tau^m} |\hat{B}_\lambda(u) - B_\lambda(u)| \\
&\leq \sup_{u\in K_\tau^m} | \sup_{c\in K_\lambda^m} E[\ell(\hat{\phi}_n'(\hat{\mathbb{G}}_n^* + u + c) - \hat{\phi}_n'(c))|\{X_i\}] - \sup_{c\in K_\lambda^m} E[\ell(\hat{\phi}_n'(\mathbb{G}_0 + u + c) - \hat{\phi}_n'(c))|\{X_i\}]| \\
&\quad + \sup_{u\in K_\tau^m} | \sup_{c\in K_\lambda^m} E[\ell(\hat{\phi}_n'(\mathbb{G}_0 + u + c) - \hat{\phi}_n'(c))|\{X_i\}] - \sup_{c\in K_\lambda^m} E[\ell(\phi_{\theta_0}'(\mathbb{G}_0 + u + c) - \phi_{\theta_0}'(c))]| \ .
\end{aligned} \tag{A.36}$$

For the first term on the right hand side, we have by Assumptions 3.3 and 3.5(iii)-(b):

$$\begin{aligned}
&\sup_{u\in K_\tau^m} | \sup_{c\in K_\lambda^m} E[\ell(\hat{\phi}_n'(\hat{\mathbb{G}}_n^* + u + c) - \hat{\phi}_n'(c))|\{X_i\}] - \sup_{c\in K_\lambda^m} E[\ell(\hat{\phi}_n'(\mathbb{G}_0 + u + c) - \hat{\phi}_n'(c))|\{X_i\}]| \\
&\leq \sup_{u\in K_\tau^m, c\in K_\lambda^m} |E[\ell(\hat{\phi}_n'(\hat{\mathbb{G}}_n^* + u + c) - \hat{\phi}_n'(c))|\{X_i\}] - E[\ell(\hat{\phi}_n'(\mathbb{G}_0 + u + c) - \hat{\phi}_n'(c))|\{X_i\}]| \\
&\leq \sup_{f\in \mathrm{BL}_a(\mathbb{D})} |E[f(\hat{\mathbb{G}}_n^*)|\{X_i\}] - E[f(\mathbb{G}_0)]| = o_p(1) \ ,
\end{aligned} \tag{A.37}$$

where $a \equiv \max\{M, C_\ell C_{\hat{\phi}'}\}$ with M being a upper bound of ℓ, and the last equality is by Assumption 3.5(i). As for the second term on the right hand side of (A.36), fix $\epsilon > 0$ and choose a compact set $K_0 \subset \mathbf{R}^m$ such that $P(\mathbb{G}_0 \notin K_0) < \epsilon/(4M)$. Then

$$\begin{aligned}
&\sup_{u\in K_\tau^m} | \sup_{c\in K_\lambda^m} E[\ell(\hat{\phi}_n'(\mathbb{G}_0 + u + c) - \hat{\phi}_n'(c))|\{X_i\}] - \sup_{c\in K_\lambda^m} E[\ell(\phi_{\theta_0}'(\mathbb{G}_0 + u + c) - \phi_{\theta_0}'(c))]| \\
&\leq \sup_{u\in K_\tau^m, c\in K_\lambda^m} E[|\ell(\hat{\phi}_n'(\mathbb{G}_0 + u + c) - \hat{\phi}_n'(c)) - \ell(\phi_{\theta_0}'(\mathbb{G}_0 + u + c) - \phi_{\theta_0}'(c))||\{X_i\}] \\
&\leq 2M \cdot P(\mathbb{G}_0 \notin K_0) + C_\ell \sup_{z\in K} \|\hat{\phi}_n'(z) - \phi_{\theta_0}'(z)\|_{\mathbb{E}} + C_\ell \sup_{z\in K_\lambda^m} \|\hat{\phi}_n'(z) - \phi_{\theta_0}'(z)\|_{\mathbb{E}} \\
&\leq \frac{\epsilon}{2} + o_p(1) \ ,
\end{aligned} \tag{A.38}$$

where $K \equiv K_0 + K_\lambda^m + K_\tau^m$ is compact, and the last step is by Lemma A.7.

Result (A.35) then follows from results (A.36), (A.37) and (A.38). The rest of the proof follows from that of Lemma (A.2). ■

Proof of Theorem 3.4: The proof is essentially the same as that of Theorem 3.2 by combining Lemmas A.4 and A.6 and is thus omitted. ■

Lemma A.4 *Suppose that Assumptions 2.1, 2.2, 2.3, 3.1, 3.2, 3.3, 3.4, 3.5, 3.6(i)(ii), and 3.8 hold. Assume that the loss function ℓ is bounded. If $\hat{u}_{n,m} \in \hat{\Psi}_m$, then for each $m \in \mathbf{N}$,*

$$d(\hat{u}_{n,m}, \Psi_m) \xrightarrow{p} 0 \ , \tag{A.39}$$

under $\{P_{n,h}\}$ for each $h \in H$.

Proof: Fix $m \in \mathbf{N}$ throughout. The proof closely follows that of Lemma A.2. First, by the same arguments as before we can show the following uniform convergence result:

$$\sup_{u\in \mathbb{D}_u} |\hat{B}_m(u) - B_m(u)| \xrightarrow{p} 0 \tag{A.40}$$

under $P_{n,0}$, and the identification condition – i.e. for each $\epsilon > 0$,

$$\inf_{u \in \mathbb{D}_u \backslash \Psi_m^\epsilon} B_m(u) > \inf_{u \in \mathbb{D}_u} B_m(u) \ . \tag{A.41}$$

Fix $\epsilon > 0$. Now by the identification result (A.41), there is some $\delta > 0$ such that whenever $u \in \mathbb{D}_u \backslash \Psi_m^\epsilon$ we have

$$B_m(u) - B_m(u_m^*) \geq 2\delta \ , \tag{A.42}$$

where u_m^* is any element in Ψ_m. Moreover, by Assumptions 3.3 and 3.4, $B_m(u)$ is continuous. Then by Assumption 3.8(ii) and the fact that $k_n \to \infty$ as $n \to \infty$, we may pick $u_{k_n} \in \mathbb{D}_{k_n}$ such that

$$B_m(u_{k_n}) - \delta \leq B_m(u_m^*) \ , \tag{A.43}$$

for all n sufficiently large. We now have for all n sufficiently large (so that (A.43) holds):

$$\begin{aligned}
P_{n,0}(d(\hat{u}_{n,m}, \Psi_m) > \epsilon) &= P_{n,0}(\hat{u}_{n,m} \in \mathbb{D}_u \backslash \Psi_m^\epsilon) \\
&\leq P_{n,0}(B_m(\hat{u}_{n,m}) - B_m(u_m^*) \geq 2\delta) \\
&= P_{n,0}(B_m(\hat{u}_{n,m}) - \hat{B}_m(\hat{u}_{n,m}) + \hat{B}_m(\hat{u}_{n,m}) - B_m(u_m^*) \geq 2\delta) \\
&\leq P_{n,0}(B_m(\hat{u}_{n,m}) - \hat{B}_m(\hat{u}_{n,m}) + \hat{B}_m(u_{k_n}) + \epsilon_n - B_m(u_m^*) \geq 2\delta) \\
&\leq P_{n,0}(B_m(\hat{u}_{n,m}) - \hat{B}_m(\hat{u}_{n,m}) + \hat{B}_m(u_{k_n}) + \epsilon_n - B_m(u_{k_n}) \geq \delta) \\
&\leq P_{n,0}(2 \sup_{u \in \mathbb{D}_{k_n}} |B_m(u) - \hat{B}_m(u)| + \epsilon_n \geq \delta) \\
&\leq P_{n,0}(2 \sup_{u \in \mathbb{D}_u} |B_m(u) - \hat{B}_m(u)| + \epsilon_n \geq \delta) \to 0 \ ,
\end{aligned} \tag{A.44}$$

where the second inequality is due to the definition of $\hat{u}_{n,m}$, and the third inequality is by (A.43). Result (A.39) then follows under $\{P_{n,0}\}$ by (A.44), (A.40) and $\epsilon_n = o_p(1)$ as $n \to \infty$. By Assumption 2.1(ii), $P_{n,h}$ and $P_{n,0}$ are mutually contiguous for each $h \in H$; see, for example, Example 6.5 in van der Vaart (1998). Then lemma follows from Le Cam's first lemma. ■

Lemma A.5 *Let $(\mathbb{D}, \tau)$ be a topological space and $f : \mathbb{D} \to \overline{\mathbf{R}}$ be a lower semicontinuous function. For any $A \subset \mathbb{D}$, we have*

$$\sup_{x \in A} f(x) = \sup_{x \in \overline{A}} f(x) \ ,$$

where $\overline{A}$ denotes the closure of A relative to τ.

PROOF: Only consider the nontrivial case when A is nonempty. Suppose first that $\sup_{x \in \overline{A}} f(x) = \infty$. Fix arbitrary large $M > 0$. Then there is some $x_0 \in \overline{A}$ such that $f(x_0) \geq M$. Since $x_0 \in \overline{A}$, we may pick a net $\{x_\alpha\} \subset A$ such that $x_\alpha \to x_0$ in τ. But then since f is lower semicontinuous, $\liminf_\alpha f(x_\alpha) \geq f(x_0)$. In turn, this implies that there is some α^* such that $\sup_{x \in A} f(x) \geq f(x_{\alpha^*}) > f(x_0) - 1 \geq M - 1$. Since M is arbitrary, it follows that $\sup_{x \in A} f(x) = \infty$.

Now suppose that $\sup_{x \in \overline{A}} f(x) < \infty$. Obviously, $\sup_{x \in A} f(x) \leq \sup_{x \in \overline{A}} f(x)$. To conclude, it suffices to show that for any $\epsilon > 0$,

$$\sup_{x \in \overline{A}} f(x) \leq \sup_{x \in A} f(x) + \epsilon \ . \tag{A.45}$$

First, we may pick some $x_0 \in \overline{A}$ such that $\sup_{x\in\overline{A}} f(x) \leq f(x_0) + \epsilon/2$. Next, we may choose a net $\{x_\alpha\} \subset A$ such that $x_\alpha \to x_0$ in τ. Since f is lower semicontinuous, $\liminf_\alpha f(x_\alpha) \geq f(x_0)$, implying that we may find some α^* such that $f(x_0) \leq f(x_{\alpha^*}) + \epsilon/2$. Combining previous two inequalities, we conclude that

$$\sup_{x\in\overline{A}} f(x) \leq f(x_0) + \epsilon/2 \leq f(x_{\alpha^*}) + \epsilon \leq \sup_{x\in A} f(x) + \epsilon \ ,$$

proving (A.45), and we thus establish the Lemma. ■

Lemma A.6 *Let $(\mathbb{D}, d)$ be a metric space and $K \subset \mathbb{D}$ a nonempty compact subset. Let $(\Omega_n, \mathcal{A}_n, P_n)$ be a sequence of probability spaces and $X_n : \Omega_n \to \mathbb{D}$ arbitrary maps such that $d(X_n, K) \xrightarrow{p} 0$ under $\{P_n\}$. Then for any subsequence $\{n_k\}$, there exist a further subsequence $\{n_{k_j}\}$ and some deterministic $c \in K$ such that $X_{n_{k_j}} \xrightarrow{p} c$ as $j \to \infty$.*

Proof: We proceed by contradiction. Fix a subsequence $\{n_k\}$ and suppose that for each $c \in K$ and every subsequence $\{n_{k_j}\}$, $X_{n_{k_j}} \not\xrightarrow{p} c$ as $j \to \infty$. This implies that for each $c \in K$ there exist $\epsilon_c > 0$ and $\eta_c \in (0, 1)$ such that

$$\liminf_{k\to\infty} P_{n_k}(d(X_{n_k}, c) > 2\epsilon_c) > \eta_c \ ,$$

or equivalently,

$$\limsup_{k\to\infty} P_{n_k}(d(X_{n_k}, c) < 2\epsilon_c) < 1 - \eta_c \ . \tag{A.46}$$

Next, for each $c \in K$, let $B_c(\epsilon_c) \equiv \{c' \in K : d(c', c) < \epsilon_c\}$. Since $\{B_c(\epsilon_c)\}_{c\in K}$ is an open cover of K, compactness of K implies that there exists a finite subcover $\{B_{c_j}(\epsilon_j)\}_{j=1}^{J^*}$ with $J^* < \infty$ and $\epsilon_j \equiv \epsilon_{c_j}$ that covers K. Observe that if $d(X_{n_k}, c_j) \geq 2\epsilon_j$ for all $j = 1, \ldots, J^*$, then we must have

$$d(X_{n_k}, K) \geq \epsilon_0 \ ,$$

where $\epsilon_0 \equiv \min(\epsilon_1, \ldots, \epsilon_{J^*})$. To see this, suppose $d(X_{n_k}, K) < \epsilon_0$ and $d(X_{n_k}, K) = d(X_{n_k}, c')$ for some $c' \in K$. Since $d(c', c_j) < \epsilon_j$ for some j, it follows that

$$d(X_{n_k}, c_j) \leq d(X_{n_k}, c') + d(c', c_j) < \epsilon_0 + \epsilon_j \leq 2\epsilon_j \ ,$$

a contradiction, implying that

$$P_{n_k}(d(X_{n_k}, K) \geq \epsilon_0) \geq P_{n_k}(d(X_{n_k}, c_j) \geq 2\epsilon_j, j = 1, \ldots, J^*) \ . \tag{A.47}$$

Elementary calculations then reveal that

$$\begin{aligned}
&\liminf_{k\to\infty} P_{n_k}(d(X_{n_k}, c_j) \geq 2\epsilon_j, j = 1, \ldots, J^*) \\
&\quad = 1 - \limsup_{k\to\infty} P_{n_k}(d(X_{n_k}, c_j) < 2\epsilon_j \text{ for some } j = 1, \ldots, J^*) \\
&\quad \geq 1 - \sum_{j=1}^{J^*} \limsup_{k\to\infty} P_{n_k}(d(X_{n_k}, c_j) < 2\epsilon_j) \geq 1 - \sum_{j=1}^{J^*} (1 - \eta_{c_j}) \equiv \eta_0 \ ,
\end{aligned} \tag{A.48}$$

where we may assume that $\eta_0 > 0$ by choosing η_{c_j}'s sufficiently small since we may increase each ϵ_c to make η_c arbitrarily close to 1 or $1 - \eta_c$ arbitrarily close to zero and

meanwhile J^* wouldn't increase because the radius of each open ball $B_c(\epsilon_c)$ of the open cover $\{B_c(\epsilon_c)\}_{c\in K}$ increases. Combination of (A.47) and (A.48) then yields

$$\liminf_{k\to\infty} P_{n_k}(d(X_{n_k}, K) \geq \epsilon_0) \geq \eta_0 > 0 \ ,$$

a contradiction. This completes the proof. ■

Lemma A.7 *Suppose Assumptions 2.1(ii) and 3.5(iii) hold. Then for any compact subset $K \subset \mathbb{D}$ and any $\epsilon > 0$,*

$$\sup_{I\subset_f H} \limsup_{n\to\infty} \sup_{h\in I} P_{n,h}(\sup_{z\in K} \|\hat{\phi}'_n(z) - \phi'_{\theta_0}(z)\|_{\mathbb{E}} > \epsilon) = 0 \ . \tag{A.49}$$

Proof: Fix a compact subset $K \subset \mathbb{D}$ and $\epsilon > 0$. Since K is compact, ϕ'_{θ_0} is continuous and hence uniformly continuous on K so that we may find a finite collection $\{z_j\}_{j=1}^J$ with $J < \infty$ such that $z_j \in K$ for all j and

$$\sup_{z\in K} \min_{1\leq j\leq J} \max\left\{C_{\hat{\phi}'}\|z - z_j\|_{\mathbb{D}}, \|\phi'_{\theta_0}(z) - \phi'_{\theta_0}(z_j)\|_{\mathbb{E}}\right\} < \frac{\epsilon}{3} \ .$$

This, along with Assumption 3.5(iii)-b), implies that

$$\sup_{z\in K} \|\hat{\phi}'_n(z) - \phi'_{\theta_0}(z)\|_{\mathbb{E}} \leq \max_{1\leq j\leq J} \|\hat{\phi}'_n(z_j) - \phi'_{\theta_0}(z_j)\|_{\mathbb{E}} + \frac{2}{3}\epsilon \ . \tag{A.50}$$

Fix a finite subset $I \subset H$. By Assumption 2.1(ii), $P_{n,h}$ and $P_{n,0}$ are mutually contiguous for each $h \in H$; see, for example, Example 6.5 in van der Vaart (1998). It follows from Assumption 3.5(iv)-a) and Le Cam's first lemma that

$$\limsup_{n\to\infty} \sup_{h\in I} P_{n,h}(\|\hat{\phi}'_n(z_j) - \phi'_{\theta_0}(z_j)\|_{\mathbb{E}} > \frac{\epsilon}{3}) = 0 \text{ for all } j = 1, \ldots, J \ . \tag{A.51}$$

Combining (A.51) and (A.51) we thus conclude that

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{h\in I} P_{n,h}(\sup_{z\in K} \|\hat{\phi}'_n(z) - \phi'_{\theta_0}(z)\|_{\mathbb{E}} > \epsilon) \\
&\qquad \leq \limsup_{n\to\infty} \sup_{h\in I} P_{n,h}(\max_{1\leq j\leq J} \|\hat{\phi}'_n(z_j) - \phi'_{\theta_0}(z_j)\|_{\mathbb{E}} > \frac{\epsilon}{3}) \\
&\qquad \leq \sum_{j=1}^{J} \limsup_{n\to\infty} \sup_{h\in I} P_{n,h}(\|\hat{\phi}'_n(z_j) - \phi'_{\theta_0}(z_j)\|_{\mathbb{E}} > \frac{\epsilon}{3}) = 0 \ .
\end{aligned}$$

Since this is true for each finite $I \subset H$, the lemma then follows immediately. ■

Appendix B Results for Examples 2.1 - 2.4

Example 2.1 (Best Treatment)

By Corollary 3.1, the lower bound when $\theta^{(1)} = \theta^{(2)}$ in this example becomes

$$\begin{aligned}
&\inf_{u\in\mathbf{R}^2} \sup_{c\in\mathbf{R}^2} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c)\right)\right] \\
&\quad = \inf_{u\in\mathbf{R}^2} \sup_{c\in\mathbf{R}^2} E\left[\left(\max\{\mathbb{G}_0^{(1)} + u^{(1)} + c^{(1)}, \mathbb{G}_0^{(2)} + u^{(2)} + c^{(2)}\} - \max\{c^{(1)}, c^{(2)}\}\right)^2\right] \ ,
\end{aligned}$$

where $\mathbb{G}_0 \sim \mathcal{N}(0, \sigma^2 I_2)$. Replace $u^{(1)}$ and $u^{(2)}$ with $u - \Delta_u$ and u respectively; similarly define $c - \Delta_c$ and c. Since the problem is symmetric in c_1 and c_2, we may assume that $\Delta_c \geq 0$. Then we have

$$\begin{aligned}
&\inf_{u \in \mathbf{R}^2} \sup_{c \in \mathbf{R}^2} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c)\right)\right] \\
&\qquad = \inf_{u \in \mathbf{R}, \Delta_u \in \mathbf{R}} \sup_{\Delta_c \geq 0} E\left[\left(\max(\mathbb{G}_0^{(1)} - \Delta_c - \Delta_u, \mathbb{G}_0^{(2)}) + u\right)^2\right] \\
&\qquad = \inf_{u \in \mathbf{R}, \Delta_u \in \mathbf{R}} \sup_{\Delta_c \geq \Delta_u} E\left[\left(\max(\mathbb{G}_0^{(1)} - \Delta_c, \mathbb{G}_0^{(2)}) + u\right)^2\right] .
\end{aligned}$$

Notice that for each $u \in \mathbf{R}$,

$$\sup_{\Delta_c \geq \Delta_u} E\left[\left(\max(\mathbb{G}_0^{(1)} - \Delta_c, \mathbb{G}_0^{(2)}) + u\right)^2\right]$$

is monotonically decreasing in Δ_u, whence we have by setting $\Delta_u = \infty$ that

$$\begin{aligned}
&\inf_{u \in \mathbf{R}^2} \sup_{c \in \mathbf{R}^2} E\left[\ell\left(\phi'_{\theta_0}(\mathbb{G}_0 + u + c) - \phi'_{\theta_0}(c)\right)\right] \\
&\qquad = \inf_{u \in \mathbf{R}} E\left[\left(\mathbb{G}_0^{(2)} + u\right)^2\right] = E[(\mathbb{G}_0^{(2)})^2] = \sigma^2 .
\end{aligned}$$

It is clear that the optimum is achieved at $u = (-\infty, 0)$ and $c = (-\infty, c^{(2)})$ with $c^{(2)} \in \mathbf{R}$ arbitrary. This is consistent with Example 6 in Song (2014a). By tedious but straightforward calculations one can show that the lower bound can be also achieved at $u = 0$ and $c = 0$.

EXAMPLE 2.2 (Interval Censored Outcome)

In this example, the identified region for ϑ is

$$\Theta_0 \equiv \{\vartheta \in \mathbf{R}^2 : E[Y_l|Z] \leq Z^\intercal \vartheta \leq E[Y_u|Z]\} .$$

Let's now work out $\sup_{\vartheta \in \Theta_0} \lambda^\intercal \vartheta$ for some fixed $\lambda \in \mathbf{R}^2$. We have

$$\begin{aligned}
&\sup\{\lambda^\intercal E[ZZ^\intercal]^{-1} E[ZE[Y|Z]] : E[Y_l|Z] \leq E[Y|Z] \leq E[Y_u|Z]\} \\
&\qquad = \sum_{j=-1}^{1} 1\{\lambda^\intercal E[ZZ^\intercal]^{-1} z_j \geq 0\} \lambda^\intercal E[ZZ^\intercal]^{-1} z_j E[Y_u|Z = z_j] P(Z = z_j) \\
&\qquad\quad + \sum_{j=-1}^{1} 1\{\lambda^\intercal E[ZZ^\intercal]^{-1} z_j < 0\} \lambda E[ZZ^\intercal]^{-1} z_j E[Y_l|Z = z_j] P(Z = z_j) ,
\end{aligned}$$

where $z_j = (1, j)^\intercal$ for $j \in \{-1, 0, 1\}$. Consider

$$\begin{aligned}
&1\{\lambda^\intercal E[ZZ^\intercal]^{-1} z_1 \geq 0\} \lambda^\intercal E[ZZ^\intercal]^{-1} z_1 \\
&\qquad = 1\{\lambda^{(1)} \frac{\theta^{(1)} + \theta^{(2)}}{\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2} + \lambda^{(2)} \frac{\theta^{(1)} - \theta^{(2)}}{\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2} \geq 0\} \\
&\qquad\quad \times \left[\lambda^{(1)} \frac{\theta^{(1)} + \theta^{(2)}}{\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2} + \lambda^{(2)} \frac{\theta^{(1)} - \theta^{(2)}}{\theta^{(1)} + \theta^{(2)} - (\theta^{(2)} - \theta^{(1)})^2}\right] \\
&\qquad \equiv 1\{\psi(\theta) \geq 0\} \psi(\theta) .
\end{aligned}$$

By the chain rule for Hadamard directionally differentiable maps (see Remark 3.1), one can show that

$$\phi'_\theta(z) = \psi'_\theta(z)1\{\psi(\theta) > 0\} + \max\{\psi'_\theta(z), 0\}1\{\psi(\theta) = 0\} \ .$$

EXAMPLE 2.3 (Incomplete Auction Model)

Lemma B.1 *Let $\phi : \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R}) \to \ell^\infty(\mathbf{R})$ be given by $\phi(\theta) = \max(\theta^{(1)}, \theta^{(2)})$, and $B_i = 1\{x : \theta_i(x) > \theta_{-i}(x)\}$ for $i = 1, 2$ and $B_0 = \{x : \theta_1(x) = \theta_2(x)\}$. It follows that ϕ is Hadamard directionally differentiable at any $\theta \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ such that for any $z \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$,*

$$\phi'_\theta(z) = z^{(1)}1_{B_1} + z^{(2)}1_{B_2} + \max\{z^{(1)}, z^{(2)}\}1_{B_0} \ .$$

PROOF: Fix $z \in \ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ and let $\{z_n\} \equiv \{(z_{1n}, z_{2n})\}$ be any sequence in $\ell^\infty(\mathbf{R}) \times \ell^\infty(\mathbf{R})$ such that $z_n \to z$ relative to the product norm as $n \to \infty$. Take arbitrary sequence $t_n \to 0$ as $n \to \infty$. Write

$$\begin{aligned}
&t_n^{-1}[\phi(\theta + t_n z_n)(x) - \phi(\theta)(x)] \\
&\quad = t_n^{-1}\left[\max\{\theta^{(1)}(x) + t_n z_{1n}(x), \theta^{(2)}(x) + t_n z_{2n}(x)\} - \max\{\theta^{(1)}(x), \theta^{(2)}(x)\}\right] \\
&\quad = t_n^{-1}\max\{t_n z_{1n}(x), \theta^{(2)}(x) - \theta^{(1)}(x) + t_n z_{2n}(x)\}1_{B_1}(x) + \max\{z_{1n}(x), z_{2n}(x)\}1_{B_0}(x) \\
&\qquad + t_n^{-1}\max\{\theta^{(1)}(x) - \theta^{(2)}(x) + t_n z_{1n}(x), t_n z_{2n}(x)\}1_{B_2}(x) \ .
\end{aligned}$$

Consider the first term. Since $t_n = o(1)$ and $z_{1n} = z_{2n} = O(1)$, for all n sufficiently large we must have

$$\max\{t_n z_{1n}(x), \theta^{(2)}(x) - \theta^{(1)}(x) + t_n z_{2n}(x)\}1_{B_1}(x) = t_n z_{1n}(x)1_{B_1}(x)$$

uniformly in $x \in \mathbf{R}$, imply that

$$t_n^{-1}\max\{t_n z_{1n}(x), \theta^{(2)}(x) - \theta^{(1)}(x) + t_n z_{2n}(x)\}1_{B_1}(x) \to z^{(1)}1_{B_1}(x)$$

uniformly in x. The third term can be handled similarly while the second term is immediate. ■

Lemma B.2 *Suppose that $\mathbb{H}$ is a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. Let $\{h_j\}_{j=1}^\infty$ be a complete sequence in $\mathbb{H}$ and $\mathbb{M} \subset \mathbb{H}$ a closed subspace. Let Π be the orthogonal projection onto $\mathbb{M}$. Then $\{\Pi h_j\}_{j=1}^\infty$ is complete in $\mathbb{M}$.*

PROOF: Fix $\epsilon > 0$ and $h \in \mathbb{M}$. Then by completeness of $\{h_j\}$ in $\mathbb{H}$, there exists $\lambda_1, \ldots, \lambda_n$ such that $\|h - (\lambda_1 h_1 + \cdots + \lambda_n h_n)\| < \epsilon$. It follows that

$$\begin{aligned}
\|h - (\lambda_1 \Pi h_1 + \cdots + \lambda_n \Pi h_n)\| &= \|\Pi h - (\lambda_1 \Pi h_1 + \cdots + \lambda_n \Pi h_n)\| \\
&= \|\Pi(h - (\lambda_1 h_1 + \cdots + \lambda_n h_n))\| \leq \|\Pi\|_{op}\|h - (\lambda_1 h_1 + \cdots + \lambda_n h_n)\| < \epsilon \ ,
\end{aligned}$$

where the second inequality follows from $\|\Pi\|_{op} = 1$ by Conway (1990, Proposition 3.3). We thus establish the Lemma.

EXAMPLE 2.4 (Quantile Curves without Crossing)

Let $\mathbb{D} = L^2(\mathcal{T}, \nu)$ where $\mathcal{T} = [\epsilon, 1-\epsilon]$ with $0 < \epsilon < 1/2$ and ν the Lebesgue measure on $\mathcal{T}$. The set Λ of (weakly) increasing functions in $\mathbb{D}$ can be formalized as follows. As standard in L^p spaces, we consider two functions in $L^2(\mathcal{T})$ to define the same element when they are equal almost everywhere. We therefore say that $f \in L^2(\mathcal{T})$ is ν-monotone or simply monotone, if there exists a monotonic function $g : \mathcal{T} \to \mathbf{R}$ such that

$$\nu(\{t \in \mathcal{T} : f(t) \neq g(t)\}) = 0 \ .$$

We then define Λ to be the set of ν-monotone functions in $L^2(\mathcal{T})$. We first show that Λ is closed and convex so that the metric projection exists and is singleton valued.

Lemma B.3 *Let $\Lambda \subset L^2(\mathcal{T})$ be the set of increasing functions. Then Λ is convex and closed.*

PROOF: Suppose that $f_1, f_2 \in \Lambda$. Then there exist increasing functions g_1 and g_2 such that $f_i = g_i$ almost everywhere. Since for any $a \in [0,1]$, $af_1 + (1-a)f_2 = ag_1 + (1-a)g_2$ almost everywhere, and $ag_1 + (1-a)g_2$ is clearly increasing, we thus conclude that $af_1 + (1-a)f_2 \in \Lambda$ and thus Λ is convex. Now take a sequence $\{f_n\} \subset \Lambda$ such that $\|f_n - f\|_{L^2} \to 0$ as $n \to \infty$. By passing to a subsequence if necessary, we may assume that $f_n \to f \in L^2(\mathcal{T})$ almost everywhere as $n \to \infty$. Since $f_n \in \Lambda$, there is an increasing function $g_n \in \Lambda$ such that $g_n = f_n$ almost everywhere. It follows that $g_n \to f$ almost everywhere. Next define, for each $t \in \mathcal{T}$,

$$\overline{f}(t) \equiv \limsup_{n \to \infty} g_n(t) \ .$$

Then $\overline{f} = f$ almost everywhere. Pick any $s, t \in \mathcal{T}$ with $s < t$, we have

$$\overline{f}(s) = \limsup_{n \to \infty} g_n(s) \leq \limsup_{n \to \infty} g_n(t) = \overline{f}(t) \ .$$

Thus f is increasing, implying that $f \in \Lambda$ and hence Λ is closed. ■

We note that if $f \in L^2(\mathcal{T})$ is monotonically increasing on $\mathcal{T}$ except a Lebesgue null set say $E_0 \subset \mathcal{T}$, then there must exist an $\tilde{f}$ such that $\tilde{f} = f$ almost everywhere and $\tilde{f}$ is increasing everywhere on $\mathcal{T}$, meaning that $f \in \Lambda$. Specifically, we construct $\tilde{f}$ as follows:

$$\tilde{f}(t) \equiv \begin{cases} f(t) & \text{if } t \in E_1 \\ \lim_{n \to \infty} f(t_n) & \text{if } t \in E_0 \end{cases} ,$$

where $E_1 \equiv \mathcal{T} \backslash E_0$, and $\{t_n\} \subset E_1$ is any sequence satisfying $t_n \downarrow t$ as $n \to \infty$. Such a sequence exists because otherwise there exists a ball $B_t(r) \equiv \{t' \in \mathcal{T} : |t' - t| \leq r\}$ for some $r > 0$ such that $B_t(r) \cap E_1 = \emptyset$ and hence $B_t(r) \subset E_0$, which is impossible since then $\nu(E_0) \geq \nu(B_t(r)) > 0$. Now it is straightforward to verify that $\tilde{f}$ is increasing on the whole domain $\mathcal{T}$. One important implication out of this is that if $f \notin \Lambda$, then there exists a Lebesgue measurable set E with $\nu(E) > 0$ such that $f(s) > f(t)$ whenever $s, t \in E$ satisfy $s < t$.

We next proceed to establish the directional differentiability of metric projection onto Λ at nonboundary points. There are couple of sufficient regularity conditions in the literature towards this end. In present case, working with polyhedricity (Haraux, 1977) is easier for us.

Lemma B.4 *Let $\mathbb{D} = L^2(\mathcal{T})$ and Λ the set of (weakly) increasing functions in $\mathbb{D}$. Then the projection Π_Λ is Hadamard directionally differentiable at any $\theta \in \mathbb{D}$ and the resulting derivative evaluated at $z \in \mathbb{D}$ is given by $\Pi_{C_\theta}(z)$, where*

$$C_\theta = T_\theta \cap [\theta - \Pi_\Lambda \theta]^\perp \ .$$

PROOF: By Haraux (1977), it suffices to show that Λ is polyhedric – i.e.

$$\overline{(\Lambda + [\Pi_\Lambda \theta]) \cap [\theta - \Pi_\Lambda \theta]^\perp} = \overline{\Lambda + [\Pi_\Lambda \theta]} \cap [\theta - \Pi_\Lambda \theta]^\perp \ . \tag{B.1}$$

In turn, polyhedricity (B.1) is immediate if we can show that $\Lambda + [\Pi_\Lambda \theta]$ is closed. To this end, consider a sequence $\{f_n\} \subset \Lambda + [\Pi_\Lambda \theta]$ such that $\|f_n - f\|_{L^2} \to 0$ for some $f \in L^2(\mathcal{T})$. We want to show that $f \in \Lambda + [\Pi_\Lambda \theta]$.

Let $\bar{\theta} = \Pi_\Lambda \theta$. Without loss of generality we may assume that $f_n = \lambda_n + a_n \bar{\theta}$ where $\lambda_n \in \Lambda$ is an increasing function for each $n \in \mathbf{N}$. If $\{a_n\}$ is bounded, then by passing to a subsequence if necessary we may assume that $a_n \to a \in \mathbf{R}$ as $n \to \infty$. This implies that $\lambda_n = f_n - a_n \bar{\theta} \to \lambda \equiv f - a\bar{\theta}$ in L^2 as $n \to \infty$. Since Λ is closed, we have $\lambda \in \Lambda$ and hence $f = \lambda + a\bar{\theta} \in \Lambda + [\Pi_\Lambda \theta]$. For unbounded $\{a_n\}$, by passing to a subsequence if necessary, first consider the case when $a_n \uparrow \infty$ with $a_n > 0$ for all $n \in \mathbf{N}$. Then $f_n = \lambda_n + a_n \bar{\theta} \in \Lambda$ for each $n \in \mathbf{N}$ since Λ is a convex cone. This immediately implies that $f \in \Lambda$ since Λ is closed and hence $f \in \Lambda + [\Pi_\Lambda \theta]$.

It remains to consider the case where $f_n = \lambda_n - a_n \bar{\theta}$ where $a_n \uparrow \infty$ and $a_n > 0$ for all $n \in \mathbf{N}$. Suppose that $f \notin \Lambda + [\Pi_\Lambda \theta]$. Then $f + a\bar{\theta}$ is not increasing for all $a \in \mathbf{R}$. In particular, $f + a_n \bar{\theta} + a\bar{\theta}$ is not increasing for each n and $a > 0$ – i.e. for each $n \in \mathbf{N}$, there is a subset $E_n \subset \mathcal{T}$ with $\nu(E_n) > 0$ such that for all $s, t \in E_n$ with $s < t$ we have

$$f(s) + a_n \bar{\theta}(s) + a\bar{\theta}(s) > f(t) + a_n \bar{\theta}(t) + a\bar{\theta}(t) \ . \tag{B.2}$$

Since $\|f_n - f\|_{L^2} \to 0$, by passing to a subsequence if necessary we may assume that $f_n \to f$ almost everywhere on $\mathcal{T}$ as $n \to \infty$. By Egoroff's theorem (Saks, 1937, p.19), we may write $\mathcal{T} = \bigcup_{j=0}^{\infty} F_j$ where $F_0, F_1, F_2, \ldots$ are Lebesgue measurable sets such that $\nu(F_0) = 0$, and $f_n \to f$ uniformly on each F_j for $j = 1, 2, \ldots$. Let $\tilde{E}_n = E_n \backslash F_0$ for all $n \in \mathbf{N}$. We claim that $\tilde{E}_n \supset \tilde{E}_{n+1}$ for each $n \in \mathbf{N}$. To see this, pick $s, t \in \tilde{E}_{n+1}$ with $s < t$ such that

$$f(s) + a_{n+1} \bar{\theta}(s) + a\bar{\theta}(s) > f(t) + a_{n+1} \bar{\theta}(t) + a\bar{\theta}(t) \ . \tag{B.3}$$

It follows that

$$\begin{aligned} f(s) + a_n \bar{\theta}(s) + a\bar{\theta}(s) &= f(s) + a_{n+1} \bar{\theta}(s) + a\bar{\theta}(s) + (a_n - a_{n+1})\bar{\theta}(s) \\ &> f(t) + a_{n+1} \bar{\theta}(t) + a\bar{\theta}(t) + (a_n - a_{n+1})\bar{\theta}(s) \\ &\geq f(t) + a_{n+1} \bar{\theta}(t) + a\bar{\theta}(t) + (a_n - a_{n+1})\bar{\theta}(t) \\ &= f(t) + a_n \bar{\theta}(t) + a\bar{\theta}(t) \ , \end{aligned}$$

where the first inequality is by (B.2), and the second is due to the facts that $a_n \leq a_{n+1}$ and that $\bar{\theta}(s) < \bar{\theta}(t)$ by $\bar{\theta} \in \Lambda$. Clearly, $f_n \to f$ everywhere as $n \to \infty$ on $\tilde{E}_1$.

To begin with, note that if there exist $s, t \in \tilde{E}_n$ with $s < t$ for some $n \in \mathbf{N}$ such that $\bar{\theta}(s) = \bar{\theta}(t)$, then by (B.2) it must be the case that $f(s) > f(t)$. Since $f_n + \alpha_n \bar{\theta}$

is monotonically increasing, $s < t$ and $\bar{\theta}(s) = \bar{\theta}(t)$, it follows that $f_n(s) \leq f_n(t)$ for all $n \in \mathbf{N}$ and hence $f(s) \leq f(t)$ by letting $n \to \infty$, a contradiction. Therefore, we may assume without loss of generality that $\bar{\theta}(s) < \bar{\theta}(t)$ for any $s, t \in \tilde{E}_1$ with $s < t$.

We further claim that $\nu(\tilde{E}_n) \downarrow 0$ as $n \to \infty$. To see this, pick $s_1, t_1 \in \tilde{E}_1$ with $s_1 < t_1$ such that (B.2) holds. We then have

$$\begin{aligned}[f(s_1) + a_n\bar{\theta}(s_1) + a\bar{\theta}(s_1)] - [f(t_1) + a_n\bar{\theta}(t_1) + a\bar{\theta}(t_1)] \\ = [f(s_1) + a\bar{\theta}(s_1)] - [f(t_1) + a\bar{\theta}(t_1)] + a_n[\bar{\theta}(s_1) - \bar{\theta}(t_1)] \\ \to -\infty < 0 \ , \quad \text{as } \ n \to \infty \ .\end{aligned}$$

Thus one of $\{s_1, t_1\}$ is not in $\bigcap_{n=1}^{\infty} \tilde{E}_n$. Continuing in this fashion, we end up with $\bigcap_{n=1}^{\infty} \tilde{E}_n$ consisting of a singleton and hence $\nu(\tilde{E}_n) \downarrow 0$ as $n \to \infty$. Since $\tilde{E}_1 \subset \bigcup_{j=1}^{\infty} F_j$, it follows that $\tilde{E}_{n_0} \subset \bigcup_{j=1}^{J} F_j$ for some n_0 and J large enough, implying that $f_n \to f$ uniformly on $\tilde{E}_{n_0}$. Thus, for all sufficiently large $n \geq n_0$ where n_0 doesn't depend on a,

$$f_n(s) + \epsilon + a_n\bar{\theta}(s) + a\bar{\theta}(s) > f_n(t) - \epsilon + a_n\bar{\theta}(t) + a\bar{\theta}(t) \ ,$$

or,

$$\lambda_n(s) + 2\epsilon + a[\bar{\theta}(s) - \bar{\theta}(t)] > \lambda_n(t) \ , \tag{B.4}$$

for $s, t \in E_n \backslash F_0$ with $s < t$. Since $\bar{\theta}(s) - \bar{\theta}(t) < 0$, by choosing $a > 0$ such that $2\epsilon + a[\bar{\theta}(s) - \bar{\theta}(t)] = 0$, we may conclude that $\lambda_n(s) > \lambda_n(t)$ for all sufficiently large $n \geq n_0$, reaching a contradiction. Hence we must have $f \in \Lambda + [\Pi_\Lambda \theta]$, meaning that $\Lambda + [\Pi_\Lambda \theta]$ is closed so that Λ is polyhedric. ■

References

Aliprantis, C. and Border, K. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. 3rd ed. Springer Verlag.

Angrist, J., Chernozhukov, V. and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, **74** 539–563.

Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, **80** pp. 313–336.

Averbukh, V. I. and Smolyanov, O. G. (1967). The theory of differentiation in linear topological spaces. *Russian Mathematical Surveys*, **22** 201–258.

Averbukh, V. I. and Smolyanov, O. G. (1968). The various definitions of the derivative in linear topological spaces. *Russian Mathematical Surveys*, **23** 67.

Bassett, J., Gilbert and Koenker, R. (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, **77** pp. 407–415.

Begun, J. M., Hall, W. J., Huang, W. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, **11** pp. 432–452.

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

Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.

Blackwell, D. (1951). Comparison of experiments. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. 93–102.

Blumenthal, S. and Cohen, A. (1968a). Estimation of the larger of two normal means. *Journal of the American Statistical Association*, **63** pp. 861–876.

Blumenthal, S. and Cohen, A. (1968b). Estimation of the larger translation parameter. *The Annals of Mathematical Statistics*, **39** pp. 502–516.

Bogachev, V. I. (1998). *Gaussian Measures*. 62, American Mathematical Society, Providence.

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

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, **34** 305 – 334.

Chandrasekhar, A., Chernozhukov, V., Molinari, F. and Schrimpf, P. (2012). Inference for best linear approximations to set identified functions. *ArXiv e-prints*.

Chernoff, H. (1956). Large-sample theory: Parametric case. *The Annals of Mathematical Statistics*, **27** pp. 1–22.

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, **78** 1093–1125.

Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, **73** 245–261.

Chernozhukov, V. and Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, **132** 491 – 525.

Chernozhukov, V., Lee, S. and Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica*, **81** 667–737.

Conway, J. (1990). *A Course in Functional Analysis*. 2nd ed. Springer.

Davydov, Y. A., Lifshits, M. and Smorodina, N. (1998). *Local Properties of Distributions of Stochastic Functionals*. American Mathematical Society.

Doss, H. and Sethuraman, J. (1989). The price of bias reduction when there is no unbiased estimate. *The Annals of Statistics*, **17** pp. 440–442.

Dudley, R. (1966). Convergence of Baire measures. *Studia Mathematica*, **27** 251–268.

Dudley, R. M. (1968). Distances of probability measures and random variables. *The Annals of Mathematical Statistics*, **39** 1563–1572.

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

Dvoretzky, A., Wald, A. and Wolfowitz, J. (1950). Elimination of randomization in certain problems of statistics and of the theory of games. *Proceedings of the National Academy of Sciences of the United States of America*, **36** pp. 256–260.

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*, **22** pp. 1–21.

Fang, Z. and Santos, A. (2014). Inference on directionally differentiable functions. Working paper, University of California, San Diego.

Feinberg, E. and Piunovskiy, A. (2006). On the Dvoretzky-Wald-Wolfowitz theorem on nonrandomized statistical decisions. *Theory of Probability & Its Applications*, **50** 463–466.

Folland, G. (1999). *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. Wiley & Sons.

Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, **55** pp. 363–390.

Haile, P. A. and Tamer, E. (2003). Inference with an incomplete model of English auctions. *Journal of Political Economy*, **111** pp. 1–51.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **14** 323–330.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, 175–194.

Haraux, A. (1977). How to differentiate the projection on a convex set in Hilbert space. some applications to variational inequalities. *Journal of the Mathematical Society of Japan*, **29** 615–631.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, **51** pp. 186–192.

Hirano, K. and Porter, J. (2012). Impossibility results for nondifferentiable functionals. *Econometrica*, **80** 1769–1790.

Hong, H. and Li, J. (2014). The numerical directional Delta method. Working paper, Stanford University.

Huber, P. J. (1966). Strict efficiency excludes superefficiency. *The Annals of Mathematical Statistics*, **37** 1425.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). Statistical estimation: Asymptotic theory.

Jeganathan, P. (1981). On a decomposition of the limit distribution of a sequence of estimators. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **43** pp. 26–36.

Jeganathan, P. (1982). On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **44** pp. 173–212.

Kaido, H. (2013). Asymptotically efficient estimation of weighted average derivatives with an interval censored variable. *ArXiv e-prints*.

Kaido, H. and Santos, A. (2013). Asymptotically efficient estimation of models defined by convex moment inequalities.

Kline, P. and Santos, A. (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure. *Quantitative Economics*, **4** 231–267.

Koshevnik, Y. and Levit, B. (1976). On a non-parametric analogue of the information matrix. *Theory of Probability & Its Applications*, **21** 738–753.

Le Cam, L. M. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. **1**.

Le Cam, L. M. (1955). An extension of Wald's theory of statistical decision functions. *The Annals of Mathematical Statistics*, **26** pp. 69–81.

Le Cam, L. M. (1964). Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, **35** 1419–1455.

Le Cam, L. M. (1972). Limits of experiments. In *Proc. Sixth Berkeley Symp. Math. Statist. Probab*, vol. 1. University of California Press, 245–261.

Le Cam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag New York.

Lee, Y. (2009). Efficiency bounds for semiparametric estimation of quantile regression under misspecification. Working paper, University of Wisconsin-Madison.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. 2nd ed. Springer.

Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer Verlag.

Levit, B. (1978). Infinite-dimensional informational bounds. *Theory of Probability and Its Applications*, **23** 371–377.

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, **68** 997–1010.

Manski, C. F. and Pepper, J. V. (2009). More on monotone instrumental variables. *Econometrics Journal*, **12** S200–S216.

Manski, C. F. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, **70** pp. 519–546.

Millar, P. (1985). Non-parametric applications of an infinite dimensional convolution theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **68** 545–556.

Millar, P. W. (1983). The minimax principle in asymptotic statistical theory. vol. 976 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, 75–265.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65** 331–355.

Newey, W. (1993). Efficient estimation of models with conditional moment restrictions. In *Econometrics* (H. G.S.Maddala, C.R.Rao, ed.), vol. 11 of *Handbook of Statistics*, chap. 16. Amsterdam: North-Holland, 419 – 454.

Newey, W. and Powell, J. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, **71** pp. 1565–1578.

Oosterhoff, J. and Zwet, W. (1979). A note on contiguity and Hellinger distance. In *Contributions to Statistics: Jaroslav Hájek Memorial Volume* (J. Jurecková, ed.). Reidel, Dordrecht, 157–166.

Pfanzagl, J. and Wefelmeyer, W. (1982). *Contributions to a General Asymptotic Statistical Theory*, vol. 13. Springer-Verlag New York.

Saks, S. (1937). *Theory of the Integral*. Courier Dover Publications.

Santos, A. (2012). Inference in nonparametric instrumental variables with partial identification. *Econometrica*, **80** 213–275.

Severini, T. A. and Tripathi, G. (2001). A simplified approach to computing bounds in semiparametric models. *Journal of Econometrics*, **102** 23 – 66.

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

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

Song, K. (2014a). Local asymptotic minimax estimation of nonregular parameters with translation-scale equivariant maps. *Journal of Multivariate Analysis*, **125** 136 – 158.

Song, K. (2014b). Minimax estimation of nonregular parameters and discontinuity in minimax risk. *ArXiv e-prints*.

Strasser, H. (1982). Local asymptotic minimax properties of Pitman estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **60** 223–247.

van der Vaart, A. W. (1988a). Estimating a real parameter in a class of semiparametric models. *The Annals of Statistics*, **16** pp. 1450–1474.

van der Vaart, A. W. (1988b). *Statistical Estimation in Large Parameter Spaces*. CWI Tracts 44, Centrum voor Wiskunde en Informatica, Amsterdam.

van der Vaart, A. W. (1989). On the asymptotic information bound. *The Annals of Statistics*, **17** pp. 1487–1500.

van der Vaart, A. W. (1991a). An asymptotic representation theorem. *International Statistical Review*, **59** pp. 97–121.

van der Vaart, A. W. (1991b). Efficiency and Hadamard differentiability. *Scandinavian Journal of Statistics*, **18** pp. 63–75.

van der Vaart, A. W. (1992). Asymptotic linearity of minimax estimators. *Statistica Neerlandica*, **46** 179–194.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1990). Prohorov and continuous mapping theorems in the Hoffmann-Jørgensen weak convergence theory, with application to convolution and asymptotic minimax theorems. Tech. Rep. 157, Department of Statistics, University of Washington, Seattle.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Verlag.

Wald, A. (1950). *Statistical Decision Functions*. Wiley.

Zarantonello, E. H. (1971). Projections on convex sets and Hilbert spaces and spectral theory. In *Contributions to Nonlinear Functional Analysis* (E. H. Zaranotello, ed.). Academic Press.