Efficient Estimation of Directionally Differentiable Functionals

Kirill Ponomarev[*]
PhD Candidate, Department of Economics, UCLA
ponomkirill@gmail.com

This version: January 3, 2022
(Check out the most recent version)

Abstract

This paper studies estimation of parameters of the form $\phi(\theta_0)$, where ϕ is a known directionally differentiable function, and θ_0 is an estimable feature of the observed distribution of the data. Such parameters are abundant in econometric models and typically take the form of maxima or minima of some estimable objects. Examples include bounds on the average treatment effects in non-experimental settings, identified sets for the coefficients in regression models with interval-valued data, bounds on the distribution of wages accounting for selection into employment, and many others. I show that the efficient (Locally Asymptotically Minimax) estimators for such parameters take the form $\phi(\hat{\theta}_n + \hat{v}_{1,n}) + \hat{v}_{2,n}$, where $\hat{\theta}_n$ is the efficient estimator for θ_0, and $\hat{v}_{1,n}, \hat{v}_{2,n}$ are suitable adjustment terms. I demonstrate that the optimal adjustment terms depend on the chosen loss function and develop a general procedure to compute them from the data. A simulation study shows that the proposed estimator can have lower finite-sample bias and variance than the existing alternatives. As an application, I construct efficient estimators for the bounds on the distribution of valuations and the optimal reserve price in English auctions with independent private values. Empirically calibrated simulations show that the resulting estimates are substantially sharper than the previously available ones.

[*] I am deeply grateful to my advisors Rosa Matzkin and Andres Santos, for their constant support and guidance. I would also like to thank Denis Chetverikov for his support and suggestions that helped to improve this paper.

1 Introduction

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

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

In this paper, I develop efficient estimators for such parameters in a general setting. Specifically, I assume that the parameter θ_0 is "well-behaved," in the sense that

[1] Other examples include bounds on structural parameters in market entry and discrete choice models (Ciliberto and Tamer, 2009; Beresteanu, Molchanov, and Molinari, 2011; Pakes, Porter, Ho, and Ishii, 2007, 2015), shape restrictions via projections (Fang, 2018), and the breakdown frontiers in the recent literature on sensitivity analysis (Kline and Santos, 2013; Masten and Poirier, 2020). A more detailed discussion is provided in Section 2.2.

a regular efficient estimator $\hat{\theta}_n$ is available, and that the function ϕ is everywhere directionally differentiable. To accommodate applications such as English auctions or regressions with interval-valued data, I allow both θ_0 and $\phi(\theta_0)$ to take values in finite or infinite-dimensional spaces. I show that an efficient estimator for $\phi(\theta_0)$ can be constructed as

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

where $\hat{v}_{1,n}$, $\hat{v}_{2,n}$ are adjustment terms, computed from the data.

The proposed estimator has two key features. First, it automatically adapts to the presence or failure of full differentiability. That is, if the data suggest that the function ϕ is likely to be fully differentiable at θ_0, both adjustment terms will be equal to zero by construction. In this case, the proposed estimator reduces to $\phi(\hat{\theta}_n)$, which is known to be efficient under full differentiability (e.g. van der Vaart, 1988). On the other hand, if the data reveal that the full differentiability is likely to fail at θ_0, the adjustment terms will differ from zero and improve on the "plug-in" estimator. Second, the optimal adjustment terms depend on the loss function chosen to evaluate and compare different estimators. Under full differentiability, the "plug-in" estimator $\phi(\hat{\theta}_n)$ is known to be efficient for any symmetric "bowl-shaped" loss function, so that the choice of a particular functional form is irrelevant (e.g. van der Vaart, 1988). In contrast, when differentiability fails, the adjustment terms can depend on the loss function, suggesting that the latter should be tailored to specific applications. In particular, choosing the squared loss function allows to select the adjustment terms to balance the bias-variance trade-off.

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

[2] It is worth-noting that, due to the potential lack of full differentiability, regular or unbiased estimators may not exist (van der Vaart, 1991; Hirano and Porter, 2012), and therefore traditional optimality criteria, searching for the "best regular" or "best minimum-variance unbiased" estimators, are inapplicable. Local Asymptotic Minimaxity is applicable more broadly, see Section 3.

the expectation is calculated with respect to the data distributed according to P.[3] For every fixed n, it is understood that the lower the risk, the better the estimator. The idea of Local Asymptotic Minimaxity is to compare estimators in terms of the asymptotic risk in a locally-worst-case scenario, that is,

$$\liminf_{n\to\infty} \sup_{\tilde{P}\in V_n(P)} \mathbb{E}_{\tilde{P}} \left\{ l\left(\sqrt{n}(\hat{\phi}_n - \phi(\theta(\tilde{P}))) \right) \right\}, \tag{2}$$

where $V_n(P) \subset \mathbf{P}$ denote certain "local neighborhoods" of P that shrink as n approaches infinity and only contain distributions that are hard to distinguish from P empirically. Any estimator sequence $\{\hat{\phi}_n\}$ that minimizes the above expression is called Locally Asymptotically Minimax (or LAM). A more precise formulation requires substantial background and is discussed in Section 3.

To obtain the LAM estimator, I proceed in two steps. First, I show that the LAM risk, given by (2), of any estimator satisfying mild regularity conditions is bounded from below by

$$\inf_{v_1,v_2} \sup_{s\in S(Z)} \mathbb{E}\left\{ l\left(\phi_0'(Z+v_1+s) - \phi_0'(s) + v_2 \right) \right\}, \tag{3}$$

where a random vector (or process) Z denotes the distributional limit of the efficient estimator sequence $\hat{\theta}_n$, the set $S(Z)$ denotes its support, and the function ϕ_0' denotes the directional derivative of ϕ at θ_0. This risk lower bound holds for all symmetric "bowl-shaped" loss functions, and parallels the familiar notion of the variance lower bound, establishing a sharp limit on the quality of estimation of the parameter $\phi(\theta_0)$ under directional differentiability. Second, I show that, with the appropriate choice of the adjustment terms, the estimator in (1) attains the risk lower bound in (3) and, therefore, this estimator is Locally Asymptotically Minimax. The optimal adjustment terms $\hat{v}_{1,n}, \hat{v}_{2,n}$ solve an optimization problem, corresponding to a suitable sample analog of (3). The problem takes a min-max form with a non-convex-concave objective function and, in general, can be computationally demanding. I discuss computational heuristics that help speed up the optimization and in some cases provide an approximate closed-form solution.

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

The finite-sample performance of the proposed estimator is investigated in a simulation study. I consider a simple setting, similar to Manski and Pepper (2000), in which the identified set for some real-valued parameter of interest is given by $[\max_{j\leqslant d_1}(\theta_{1,j}),\ \min_{k\leqslant d_2}(\theta_{2,k})]$, where $(\theta_1,\theta_2)=(\mathbb{E}_P(X_1),\mathbb{E}_P(X_2))\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$ for observable random vectors (X_1,X_2). Letting $(\bar{X}_{1,n},\bar{X}_{2,n})$ denote the corresponding sample means, one can estimate the bounds by $[\max_{j\leqslant d_1}(\bar{X}_{1,j,n}),\min_{k\leqslant d_2}(\bar{X}_{2,k,n})]$. However, the resulting estimates are generally biased towards each other, and, in practice, may be significantly tighter than the population bounds, potentially leading to erroneous conclusions. Therefore, it is customary to use bias-correction methods in practice (Kreider and Pepper, 2007; Chernozhukov, Lee, and Rosen, 2013). By extensive simulations, I compare the performance of the proposed efficient estimator with the simple "plug-in" estimator and the existing bias-correction methods near the values of (θ_1,θ_2) where the finite-sample bias is most problematic. These are the values (θ_1,θ_2) where the maximum/minimum are attained by multiple coordinates of θ_1 and θ_2 respectively,[4] so that the maximum/minimum functions are not fully differentiable. With the squared loss function, I find that the proposed efficient estimator mildly reduces the bias but avoids substantial fluctuations in variance, compared to the alternatives.

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

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

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

The most closely-related papers, also studying efficient estimation under directional differentiability, are Song (2014) and Fang (2018). Both papers employ versions of the LAM criterion introduced above and find that the plug-in estimator can be improved by introducing an additive adjustment term, i.e., using an estimator $\phi(\hat{\theta}_n + \hat{v}_{1,n}/\sqrt{n})$, where $\hat{\theta}_n$ is the efficient estimator for θ_0. Song (2014) focuses on Euclidean parameters $\theta_0 \in \mathbb{R}^d$ and real-valued $\phi(\theta_0)$ for a restricted class of functions ϕ. Fang (2018) allows both θ_0 and $\phi(\theta_0)$ to take values in general normed spaces but restricts the class of competing estimators to all plug-in estimators of the form $\phi(\tilde{\theta}_n)$ where $\tilde{\theta}_n$ is an arbitrary regular estimator for θ_0. This restriction excludes some important estimators, such as Stein-type shrinkage estimators, from consideration. In this paper, I do not impose restrictive assumptions on the function ϕ or the class of competing estimators and find that, in general, two adjustment terms $\hat{v}_{1,n}$ and $\hat{v}_{2,n}$ (as in Equation 1) are necessary to attain the risk lower bound. In a special case when the directional derivative ϕ_0' is translation equivariant,[5] it is possible to find an efficient estimator with $\hat{v}_{2,n} = 0$, which, for symmetric loss functions, will coincide with the one proposed by Fang (2018). In such case, the added value of this paper is in showing that the estimator is optimal among all estimators, not only plug-in estimators with regular $\tilde{\theta}_n$. In other cases, the estimator in Fang (2018) is not optimal and can be improved by adding the appropriately chosen second adjustment term.

Another closely-related paper is Fang and Santos (2019). My work is complementary to theirs: I focus on efficient estimation, whereas they focus on valid inference

[5] That is, $\phi_0'(x + c \cdot \bar{1}) = \phi_0'(x) + c$ for any $c \in \mathbb{R}$, where $\bar{1}$ denotes a vector of ones of a suitable dimension.

in settings with directionally differentiable functions.

The rest of the paper is organized as follows. Section 2 provides the general setup and motivating examples and discusses the appropriate notion of directional differentiability. Section 3 elaborates on the optimality criterion, provides some background, and formulates the basic assumptions. Sections 4 and 5 establish the general risk lower bound under directional differentiability and construct efficient estimators. Section 6 presents a simulation study. Section 7 contains an empirical application. Section 8 discusses extensions, and Section 9 concludes.

2 Directionally Differentiable Parameters

2.1 General Setup

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

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

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

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

$\mathbb{D}$ respectively, with respect to the joint law $\prod_{i=1}^{n} P_n$ of X_1^n. The individual laws P_n may change with n.

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

2.2 Motivating Examples

Next, I present several motivating examples, some of which I revisit throughout the paper to fix ideas. These examples cover both finite and infinite-dimensional parameters and include models of treatment effects (Example 1), discrete choice (Example 2), English auctions (Example 3), regression models with interval-valued data (Example 4), and shape restrictions via projection (Example 5). To focus on the main ideas, the examples are simplified.

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

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

implies

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

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

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

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

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

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

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

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

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

Example 2 (Counterfactuals in Moment Inequality Models)**.** Suppose an agent chooses $y \in \mathbb{R}^{d_Y}$ from a set $\mathcal{Y} = \{y_1, \ldots, y_M\}$ to maximize her expected payoff $\mathbb{E}(\pi(y, Z, \gamma_0)|\mathcal{F})$, where Z is a vector of payoff-relevant variables, γ_0 is a vector of payoff parameters, and $\mathcal{F}$ is the agent's information set. Let Y^* denote the optimal choice, and assume that (Y^*, Z) are observed by the econometrician. Then, optimality of Y^* implies that for all $y' \in \mathcal{Y}$,

$$\mathbb{E}(\pi(y', Z, \gamma_0) - \pi(Y^*, Z, \gamma_0)|\mathcal{F}) \leqslant 0. \tag{4}$$

A common payoff specification is $\pi(y, Z, \gamma_0) = u(y, Z) + y^T \gamma_0$, where u is a known function (e.g., Pakes, 2010). Under suitable assumptions, the optimality condition in (4) implies that γ_0 must satisfy, for any $y, y' \in \mathcal{Y}$,

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

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

$$\Gamma_0 = \{\gamma \in \mathbb{R}^{d_\gamma} : \mathbb{E}(m_{1j}(X) + m_{2j}(X)^T \gamma) \leqslant 0, j = 1, \ldots, J\}, \tag{5}$$

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

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

where $\theta_0 \in \mathbb{R}^{2J}$ is a vector of moments containing $\mathbb{E}(m_{1j}(X))$ and $\mathbb{E}(m_{2j}(X))$ for all $j = 1, \ldots, J$ and the function $F(\theta_0, \gamma)$ defines the inequalities. In this example, $\mathbb{B} = \mathbb{R}^{2J}$, $\mathbb{D} = \mathbb{R}^2$, and the function $\phi : \mathbb{R}^{2J} \to \mathbb{R}^2$ is given by $\phi(\theta) = [L(\theta), U(\theta)]$. This function is not fully differentiable whenever the above optimization problems have multiple solutions. A conceptually different approach to identification in an overlapping class of models has been developed in Galichon and Henry (2011) and

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

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

Example 3 (English Auctions)**.** Consider a symmetric ascending auction with independent private values. Each bidder draws her valuation $V_i \in [\underline{v}, \overline{v}]$, independently of the others, from a distribution with a cumulative distribution function (CDF) denoted by F. Let B_i denote the final bid of player i. For simplicity, suppose that each auction has N bidders, and the reserve price is below $\underline{v}$. The main parameter of interest in the empirical analysis in this setting is the CDF of valuations F. To relate the unobserved valuations with the observed bids, Haile and Tamer (2003) assume that each player: (i) does not bid above her valuation and (ii) does not let the others win at a price she is willing to pay. Assumption (i) can be used to obtain an upper bound on the distribution of valuations

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

where $G_{i:N}$ is the CDF of the i-th smallest bid, and $\psi_i : [0,1] \to [0,1]$ is a strictly increasing differentiable function.[7] In turn, Assumption (ii) can be used to obtain a lower bound using the distribution of the winning bid. Let $D([\underline{v}, \overline{v}], [0,1])$ denote the set of all cádlág functions from $[\underline{v}, \overline{v}]$ to $[0,1]$ (i.e., functions that are continuous from the right and have left limits evewyehrer) endowed with the supremum norm. Focusing on the upper bound presented above, in this example, $\mathbb{B} = D([\underline{v}, \overline{v}], [0,1])^N$, $\mathbb{D} = D([\underline{v}, \overline{v}], [0,1])$, $\theta_0 = (\psi_1(G_{1:N}), \dots, \psi_N(G_{N:N})) \in \mathbb{B}$ and $\phi : \mathbb{B} \to \mathbb{D}$, is defined by

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

This function is not fully differentiable if the minimum is attained by multiple $\theta_{0,i}$ for at least one $v \in [\underline{v}, \overline{v}]$. For example, if the bids are i.i.d., all $\psi_i(G_{i:N}(v))$ will coincide

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

for all $v \in [\underline{v}, \overline{v}]$. The bounds on the distribution of valuations can be translated into the bounds on the expected revenue, bidders surplus, and optimal reserve price; see Haile and Tamer (2003). In the same setting, Chesher and Rosen (2017) characterize the sharp bounds on the distribution of valuations using tools from the theory of random sets. Aradillas-López, Gandhi, and Quint (2013) provide bounds on the expected revenue and bidders surplus in auctions with correlated private values. ■

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

Example 4 (Interval Outcome Regression)**.** Let $Y \in \mathbb{R}$ be an outcome variable, $Z \in \mathbb{R}^{dz}$ be a vector of covariates, and $\beta_0 \in \mathbb{R}^{dz}$ be a vector of coefficients for the best linear prediction

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

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

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

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

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

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

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

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

Example 5 (Quantiles without Crossing)**.** Let $Y \in \mathbb{R}$ and $Z \in \mathbb{R}^d$ denote the outcome variable and the set of covariates correspondingly, and consider the quantile regression model

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

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

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

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

2.3 Hadamard Directional Differentiability

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

Definition 2.1. *A function $\phi : \mathbb{B} \to \mathbb{D}$ is Hadamard directionally differentiable at θ_0 if there is a continuous function $\phi_0' : \mathbb{B} \to \mathbb{D}$ such that, for any $h_n \to h$ in $\mathbb{B}$, and*

[8] This provides an alternative to the monotone rearrangement operator of Chernozhukov, Fernández-Val, and Galichon (2010).

any $t_n \downarrow 0$,

$$\lim_{n\to\infty} \left|\left| \frac{\phi(\theta_0 + t_n h_n) - \phi(\theta_0)}{t_n} - \phi_0'(h) \right|\right|_{\mathbb{D}} = 0. \tag{6}$$

If the above holds for each $h \in \mathbb{B}_0 \subset \mathbb{B}$*, it is said that* ϕ *is directionally differentiable at* θ_0 *tangentially to* $\mathbb{B}_0$*. In this case, the domain of* ϕ_0' *is* $\mathbb{B}_0$*.*

Intuitively, a function is directionally differentiable at θ_0 if it can be linearly approximated in each direction around θ_0, and the approximation is suitably continuous. To compare, a function ϕ is *Hadamard fully differentiable* if the derivative ϕ_0', satisfying (6), is a continuous *linear* function. That is, full differentiability implies directional differentiability, and the only distinction between the two notions is the potential non-linearity of the directional derivative (Shapiro, 1990).

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

Assumption 2.1 (Restrictions on ϕ)**.** *The map* $\phi : \mathbb{B} \to \mathbb{D}$ *is directionally Hadamard differentiable at* θ_0 *tangentially to* $\mathbb{B}_0$*, as in Definition 2.1. Moreover, the directional derivative* $\phi_0' : \mathbb{B}_0 \to \mathbb{D}$ *is Lipchitz-continuous. That is,*

$$||\phi_0'(x) - \phi_0'(y)||_{\mathbb{D}} \leqslant C_{\phi'} \, ||x - y||_{\mathbb{B}}$$

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

Since continuous linear functions are Lipchitz-continuous, this assumption is satisfied whenever ϕ is fully differentiable. Otherwise, it only imposes a mild restriction: the directional derivative is a "partially linear" function with different "slopes" in different regions of the domain, so the assumption merely rules out unbounded "slopes". Moreover, in most applications, the function ϕ itself is Lipschitz-continuous, in which case Assumption 2.1 is automatically satisfied; see Shapiro (1990).

2.3.1 Examples Revisited

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

Example 1 (Continued)**.** Focus on the upper bound $\phi(\theta_0) = \min_{j \leqslant d}(\theta_{0,j})$ with $\theta_0 \in$

$\mathbb{R}^d$. For each $h = (h_1, \ldots, h_d)^T$, the directional derivative is equal to

$$\phi_0'(h) = \min_{j \in B(\theta_0)} (h_j), \tag{7}$$

where $B(\theta_0) = \{j : \theta_{0,j} = \min_i(\theta_{0,i})\}$ is the set of indices of the components of θ_0 that attain the minimum. That is, the function ϕ is fully differentiable at θ_0 if there is a unique minimal component, and only directionally differentiable otherwise. The directional derivative satisfies Assumption 2.1 with $C_{\phi'} = 1$. Similar arguments hold for the lower bound, and for both bounds simultaneously. ■

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

$$\begin{aligned} S_1(\theta_0) &= \{v \in [\underline{v}, \overline{v}] : \theta_{1,0}(v) < \theta_{2,0}(v)\} \\ S_2(\theta_0) &= \{v \in [\underline{v}, \overline{v}] : \theta_{2,0}(v) < \theta_{1,0}(v)\} \\ S_0(\theta_0) &= \{v \in [\underline{v}, \overline{v}] : \theta_{1,0}(v) = \theta_{2,0}(v)\}. \end{aligned} \tag{8}$$

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

$$\begin{aligned} \phi_0'(h)(v) = h_1(v) \cdot \mathbf{1}(v \in S_1(\theta_0)) + h_2(v) \cdot \mathbf{1}(v \in S_2(\theta_0)) \\ + \min\{h_1(v), h_2(v)\} \cdot \mathbf{1}(v \in S_0(\theta_0)) \end{aligned} \tag{9}$$

for any $h = (h_1, h_2) \in D([\underline{v}, \overline{v}], [0,1])^2$. Therefore, whenever the set S_0 is non-empty, the function ϕ is only directionally differentiable. For example, if the bids are i.i.d., then $\psi_i(G_{i:2}) = G$ for $i = 1, 2$, so that $S_0 = [\underline{v}, \overline{v}]$. The directional derivative satisfies Assumption 2.1 with $C_\phi' = 1$. ■

3 Local Asymptotic Minimaxity

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

3.1 General Idea

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

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

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

$$\liminf_{n\to\infty} \sup_{\tilde{P}\in V_n(P)} \mathbb{E}_{\tilde{P}} \left(l \left(\sqrt{n}(\hat{\phi}_n - \phi(\theta(\tilde{P}))) \right) \right). \tag{10}$$

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

3.2 Background and Assumptions

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

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

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

These spaces are endowed with the standard $L_2(P)$ norm $||h||_{2,P} = (\int h^2 dP)^{1/2}$ and scalar product $\langle h_1, h_2 \rangle_P = \int h_1 h_2 dP$. For any subset H, of $L_2(P)$, $\bar{H}$ denotes its closure with respect to $||\cdot||_{2,P}$. To simplify exposition, I assume that the model $\mathbf{P}$ is dominated by a positive, sigma-finite measure μ on $(\mathbf{X}, \mathcal{B})$.

3.2.1 Smooth Parametric Submodels and Tangent Sets

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

Definition 3.1 (Smooth Parametric Submodels and Scores). *A smooth parametric submodel $t \mapsto P_{t,h}$ is a mapping defined on $[0, \varepsilon)$ for some $\varepsilon > 0$, such that (i) $P_{t,h}$ is a probability distribution for each t; (ii) $P_{0,h} = P$; and (iii) for some measurable function $h : \mathbf{X} \to \mathbb{R}$,*

$$\int \left(\frac{\sqrt{p_{t,h}} - \sqrt{p}}{t} - \frac{1}{2}\sqrt{p}h \right)^2 d\mu \to 0 \quad \text{as } t \downarrow 0. \tag{11}$$

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

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

Definition 3.2 (Tangent Set)**.** *The set of all scores corresponding to the submodels $\{P_{t,h}\} \subset \mathbf{P}$ is called the tangent set and denoted by*

$$T(P) = \{h \in L_0^2(P) \mid h \textit{ satisfies } (11) \textit{ for some } \{P_{t,h}\} \subset \mathbf{P}\}. \tag{12}$$

The tangent set depends on both the distribution P and the model $\mathbf{P}$ and describes the informational content of the assumption $P \in \mathbf{P}$. It is directly related to both construction of efficient estimators (e.g., Bickel, Klaassen, Ritov, and Wellner, 1993) and existence of specification tests with non-trivial power (Chen and Santos, 2018). Assumptions on $\mathbf{P}$ may translate into further restrictions on the tangent set through the requirement $\{P_{t,h}\} \subset \mathbf{P}$. If $T(P) = L_2^0(P)$, the tangent set is said to be unrestricted; otherwise, it is restricted. In the latter case, the tangent set typically forms a linear subspace of $L_2^0(P)$, but in some cases $T(P)$ can be a convex cone, e.g., in some moment inequality models.[9]

Throughout the paper, I assume that the tangent set is a linear space, as recorded below. A partial extension of the main results to convex cones and some issues associated with such settings are discussed in Section 8.

Assumption 3.1 (Random Sampling and Restrictions on the Model)**.** *The researcher observes an i.i.d. sample $\{X_i\}_{i=1}^n$ of $X \in \mathbf{X}$ from $P \in \mathbf{P}$. The model $\mathbf{P}$ and the distribution $P \in \mathbf{P}$ are such that tangent set $T(P)$ is a linear subspace of $L_0^2(P)$.*

3.2.2 Differentiable Parameters and Regular Estimators

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

Definition 3.3 (Path-Wise Differentiable Parameters)**.** *A parameter $\theta(P) \in \mathbb{B}$ is differentiable relative to a tangent set $T(P)$ if there is a continuous linear functional*

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

$\theta_0' : \bar{T}(P) \to \mathbb{B}$, *such that*

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

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

Assumption 3.2 (Diferentiability of $\theta(P)$)**.** *The parameter $\theta(P)$ is differentiable relative to the tangent set $T(P)$, according to Definitions 3.1, 3.2, and 3.3.*

Path-wise differentiability guarantees existence of the estimators with nice asymptotic behavior. The path-wise derivative θ_0' is crucial in characterizing the asymptotic efficiency bound for $\theta(P)$, which is discussed in more details in Section 3.2.3.[10] With i.i.d. data, this assumption limits the analysis to parameters estimable at the root-n rate. Examples include moments, distribution functions, quantile functions, parametric components in semi-parametric models, and smooth functions of those. Differentiable parameters are typically estimated with regular estimators.

Definition 3.4 (Regular Estimator)**.** *A sequence of estimators $\hat{\theta}_n : X_1^n \to \mathbb{B}$ for a parameter $\theta(P) \in \mathbb{B}$ is regular, if*

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

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

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

3.2.3 Convolution Theorem and Best Regular Estimators

The efficient estimator for $\phi(\theta_0)$, developed in the sequel, relies on the notion of the best regular estimator for θ_0, discussed below. Consider estimating a differentiable parameter $\theta_0 = \theta(P)$. The Convolution Theorem states that the asymptotic distribution of *any regular estimator* $\hat{\theta}_n$ can be represented as a convolution of a centered

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

Gaussian random element $\mathbb{G}_0$ and an independent "noise term" $\mathbb{W}$, that is

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

Since convolution increases variance, the "best possible" limit among regular estimators is $\mathbb{G}_0$, and its variance-covariance matrix of $\mathbb{G}_0$ is known as the *efficiency bound*. Any regular estimator that attains this limit is called the *best regular estimator*. The covariance structure and the support of $\mathbb{G}_0$ are determined by the path-wise derivative θ_0' and the tangent set $T(P)$ (see Theorems A.4 and A.5 in the Appendix).[11]

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

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

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

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

3.3 Local Asymptotic Maximum Risk

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

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

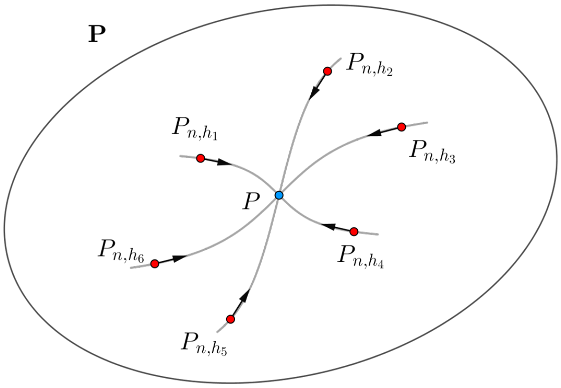


Figure 1: Example of a Local Neighborhood with $I = \{h_1, \ldots, h_6\}$.

I define the LAM risk as[12]

$$\sup_{I \subset T(P)} \liminf_{n \to \infty} \sup_{h \in I} \mathbb{E}_{P_{n,h}} \left\{ l \left(\sqrt{n}(\hat{\phi}_n - \phi(\theta(P_{n,h}))) \right) \right\}, \tag{13}$$

where I denotes an arbitrary *finite* subset $I \subset T(P)$ of the tangent set, and $P_{n,h}$ denotes a probability distribution corresponding to a smooth parametric submodel $\{P_{t,h}\} \subset \mathbf{P}$ with a score $h \in T(P)$ with $t = 1/\sqrt{n}$. In the notation of Equation (10), the local neighborhoods are $V_n(P) = \{P_{n,h} : h \in I\}$. Figure 1 illustrates.

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

3.4 Loss Functions

An essential ingredient in the LAM-analysis is the loss function. It specifies which deviations of the estimator from the estimand should be punished relatively more than the others, and by how much. In practice, the loss function can be used to "fine-tune" the estimator (e.g. specify the relative importance of different dimensions of

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

the target parameter, or focus on a subvector), address sensitivity to outliers in the data (e.g., consider the absolute loss instead of quadratic loss), or boost computation (e.g., pick a smooth or convex function). In theory, the loss function must ensure that the LAM risk is finite for at least one estimator, for otherwise the optimality criterion becomes meaningless (see, e.g., Lemma 3.1 in Fang, 2018).

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

Assumption 3.3 (Loss Functions)**.** *The loss function $l : \mathbb{D} \to \mathbb{R}_+$ is sub-convex. That is, the lower level sets $\{x \in \mathbb{D} : l(x) \leqslant c\}$ are closed, convex and symmetric.*

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

- For $x \in \mathbb{R}^d$, one can consider a weighted quadratic loss, absolute loss, or maximum loss, with $w_1, \ldots, w_d \geqslant 0$:

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

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

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

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

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

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

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

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

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

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

4 New LAM Theorem

To obtain a Locally Asymptotically Minimax estimator, I proceed in two steps. First, I derive a lower bound for the LAM risk defined in (13). This bound establishes a sharp limit on the quality of estimation of directionally differentiable parameters and suggests the form of the efficient estimator. Second, I construct an efficient estimator that attains the bound.

4.1 General Lower Bound

This section contains the first main result of the paper, which provides an extension of the LAM Theorem[13] to a class of directionally differentiable parameters. Theorem 1 below presents the general result, and Corollary 1.1 specializes to Euclidean parameters.

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

Theorem 1 (General Lower Bound)**.** *Let Assumptions 2.1, 3.1, 3.2, 3.3, and assume that the infimum in the display below can be attained. Then, for any asymptotically*

[13] Theorem 25.21 in van der Vaart (2000); Theorem 3.11.5 in van der Vaart and Wellner (1996).

tight and asymptotically measurable estimator sequence $\hat{\phi}_n : X_1^n \to \mathbb{D}$,

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

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

Corollary 1.1 (Lower Bound for Euclidean Parameters)**.** *Let Assumptions 2.1, 3.1, 3.2, 3.3, and assume that the infimum in the display below can be attained. Consider $\theta \in \mathbb{R}^{d_\theta}$ and $\phi \in \mathbb{R}^{d_\phi}$. Then, for any root-n consistent estimator sequence $\hat{\phi}_n : X_1^n \to \mathbb{R}^{d_\phi}$,*

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

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

Several comments are in order. First, note that Theorem 1 covers all reasonable estimators for $\phi(\theta_0)$. Asymptotic tightness is necessary for an estimator to converge to a tight limiting law, while asymptotic measurability is weaker than measurability for each n.[14] For Euclidean parameters, this covers all root-n consistent estimators, including, for example, Hodges-type super-efficient estimators and Stein-type shrinkage estimators (see e.g., van der Vaart, 2000; Lehmann and Casella, 2006).

[14] The discussion of measurability is only relevant in non-separable spaces. A typical estimator is a map $\hat{\phi}_n$ from $\mathbb{R}^{ndx}$ (where X_1^n lives) into the parameter space $\mathbb{D}$. Measurability would require $\hat{\phi}_n^{-1}(D)$ to be Borel in $\mathbb{R}^{ndx}$ for each Borel subset D of $\mathbb{D}$. In non-separable spaces the Borel sigma field in $\mathbb{D}$ is too rich so that many useful maps fail to be measurable. Therefore, requiring measurability for each n rules out reasonable estimators, and it is replaced with a weaker notion of asymptotic measurability; see Chapters 1.1–1.3 in van der Vaart and Wellner (1996).

Second, if the function ϕ is fully differentiable at θ_0, the lower bound simplifies as follows:

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

where (a) follows from the linearity of ϕ_0', and (b) follows from the Anderson's Lemma, since $\phi_0'(\mathbb{G}_0)$ is Gaussian. The expression $\mathbb{E}\left\{l\left(\phi_0'(\mathbb{G}_0)\right)\right\}$ is the well-known risk lower bound for differentiable parameters (e.g., van der Vaart and Wellner, 1996). It implies, in particular, that the "plug-in" estimator $\phi(\hat{\theta}_n)$ is Locally Asymptotically Minimax for any sub-convex loss function. In contrast, the lower bound in Theorem 1 suggests that with directionally differentiable functions ϕ, the optimal estimator will generally depend on the chosen loss function.

Third, the min-max form of the lower bound is not surprising. The supremum appears by construction, because the theorem deals with the locally maximum risk. In turn, the infimum appears because the lower bound must hold for a large class of competing estimators. Importantly, $v_1 \in \mathbb{B}$ and $v_2 \in \mathbb{D}$ are constants, so that the infimum is taken over $\mathbb{B} \times \mathbb{D}$, rather than over all probability distributions on this set, which makes the lower bound useful in practice. This is made possible by the purification result of Feinberg and Piunovskiy (2006) extending the seminal work of Dvoretzky, Wald, and Wolfowitz (1951) on matching randomised decision rules with nonrandomised alternatives.

Finally, the lower bound for Euclidean parameters takes a somewhat simpler form. Specifically, note that in Theorem 1, the supremum is taken over the support of $\mathbb{G}_0$, which is equal to the closure of the image of the tangent set under the path-wise derivative mapping. If the tangent set is restricted in a complicated way, this set may be hard to characterize. In contrast, the range of the efficient covariance matrix Σ_θ is a relatively simple object. In particular, if Σ_θ is of full rank, $R(\Sigma_\theta) = \mathbb{R}^{d_\theta}$.

Remark 1. To study the (LAM) estimators attaining the lower bound, it will be necessary to work with bounded loss functions, because an application of the Portmanteau lemma is required to establish the distributional convergence of the candidate estimator uniformly over the finite neighborhoods of P. To this end, let l be a loss function satisfying Assumption 3.3, and l_M be a sequence of bounded, Lipschitz-continuous sub-convex loss functions, converging to l poitwise monotonically from below. For instance, if l is continuous, one can simply take $l_M = \min\{l, M\}$ for M large enough (Lemma A.8 in the Appendix provides a general construction). Then, in the notation of Theorem 1, the lower bound also holds in the following sense:

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

4.2 Examples Revisited

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

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

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

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

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

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

Example 3 (Continued)**.** Suppose again that $N = 2$. Let $\hat{\theta}_n = (\psi_1(\hat{G}_{1:2}), \psi_2(\hat{G}_{2:2}))$ where $\hat{G}_{j:2}$, for $j = 1, 2$ are the empirical CDFs of order statistics of bids. Under suitable assumptions, it can be shown that the model $\mathbf{P}$ is unrestricted. Therefore, $\hat{G}_{1:2}, \hat{G}_{2:2}$ are best regular estimators for $G_{1:2}, G_{2:2}$, and, since ψ_1 and ψ_2 are fully Hadamard differentiable, $\hat{\theta}_n$ is the best regular estimator for θ_0. Moreover, $\sqrt{n}(\hat{\theta}_n - \theta_0)$

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

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

where the directional derivative is given in Equations (22)–(23). ■

5 Constructing LAM Estimators

Theorem 1 and Remark 1 suggest that the LAM risk of any reasonable estimator for $\phi(\theta_0)$ is bounded from below by

$$\inf_{(v_1,v_2)\in\mathbb{B}\times\mathbb{D}} \sup_{s\in S(\mathbb{G}_0)} \mathbb{E}\left\{ l_M\left(\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2\right)\right\}. \tag{14}$$

In this section, I show that the LAM estimator attaining the bound takes the form

$$\hat{\phi}_n = \phi\left(\hat{\theta}_n + \frac{\hat{v}_{1,n}}{\sqrt{n}}\right) + \frac{\hat{v}_{2,n}}{\sqrt{n}}, \tag{15}$$

where the adjustment terms $(\hat{v}_{1,n},\hat{v}_{2,n})$ converge in probability to some minimizers of (14). A natural way of obtaining such $(\hat{v}_{1,n},\hat{v}_{2,n})$ is by minimizing a suitable sample analog of (14), as discussed below.

5.1 Setup and Assumptions

Denote the population criterion function by

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

To construct a sample analog, one has to estimate two unknown components: the distribution of $\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2$ and the support $S(\mathbb{G}_0)$. The law of $\mathbb{G}_0$

can typically be approximated by bootstrap or simulation, so the main complication here is that the directional derivative ϕ_0' is an unknown and potentially non-linear function. Letting $\hat{\mathbb{G}}_n^*$ denote a bootstrap process approximating $\mathbb{G}_0$ and $\hat{\phi}_n'$ denote a suitable estimator for the directional derivative ϕ_0', the analogy principle suggests approximating the distribution of $\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2$ by the finite-sample distribution of $\hat{\phi}_n'(\hat{\mathbb{G}}_n^* + v_1 + s) - \hat{\phi}_n'(s) + v_2$ conditional on the data. Next, since $\mathbb{G}_0$ is tight, its support is separable and can be approximated by a sequence of compact sieves. It is not a substantial loss of generality to assume that $\mathbb{G}_0$ is non-degenerate, in which case the support is typically known, but more generally it has to be estimated. Let $(R_n)_{n\geqslant 1}$ denote a sequence of sieves approximating $S(\mathbb{G}_0)$ and $(\hat{R}_n)_{n\geqslant 1}$ denote the corresponding estimators. Then, I choose $(\hat{v}_{1,n}, \hat{v}_{2,n})$ to minimize:

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

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

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

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

Assumption 5.1 (Bootstrap Consistency)**.**

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

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

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

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

Condition (i) states that the limiting law of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ can be approximated by the law of $\hat{\mathbb{G}}_n^* = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$, conditional on the data.[15] Condition (ii) is a mild measurability assumption that ensures that the bootstrap process converges to $\mathbb{G}_0$ unconditionally.

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

Assumption 5.2 (Estimating the Directional Derivative)**.** *The estimator* $\hat{\phi}_n' : X_1^n \to \mathbb{D}$ *of* ϕ_0' *satisfies, for any* $\delta > 0$,

$$\sup_{s \in R_n^\delta} \left|\left|\hat{\phi}_n'(s) - \phi_0'(s)\right|\right|_{\mathbb{D}} = o_P(1),$$

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

In view of applying the extremum estimation arguments, the distribution of $\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2$ must be approximated uniformly in $(v_1, v_2) \in K$ and $s \in R_n$, where K is a fixed compact set and R_n denotes an expanding sequence of compact sets (specified in Assumption 5.3). Therefore, the estimator $\hat{\phi}_n'$ must approximate the derivative ϕ_0' uniformly well. While the above assumption may seem restrictive, natural estimators typically have a stronger property that $\hat{\phi}_n'(b) = \phi_0'(b)$ for all $b \in \mathbb{B}$ with probability approaching one. In practice, such estimators can be based on the analytical expression for ϕ_0' or obtained by numerical differentiation (see Fang and Santos, 2019; Hong and Li, 2020).

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

Assumption 5.3 (Estimating the Support)**.** *There is an expanding sequence of compact sets* $(R_n)_{n \geqslant 1} \subset \mathbb{B}$ *such that for any* $\varepsilon > 0$ *and* $s \in S(\mathbb{G}_0)$*, there is* $s_n \in R_n$ *for* n *large enough such that* $||s_n - s|| \leqslant \varepsilon$*. The sets* R_n *are either known or can be estimated with* $\hat{R}_n$ *satisfying* $d_H(\hat{R}_n, R_n) = o_P(1)$ *as* $n \to \infty$*.*

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

Recall that $S(\mathbb{G}_0)$ is equal to the closure of $\theta_0'(T(P))$ in $\mathbb{B}$. Since both θ_0' and $T(P)$ are unknown and the latter may be restricted in a non-trivial way, estimating $S(\mathbb{G}_0)$ is, in general, a complicated task. However, as I show below, Assumption 5.3 can be verified in a number of different ways, depending on the application, and does not necessarily require estimating the tangent set $T(P)$ and the path-wise derivative θ_0' directly. See Sections 5.2.1 and 5.2.2 for further discussion and examples.

Finally, note that the minimization problems with both Q and $\hat{Q}_n$ may have multiple solutions. It is therefore necessary to formulate conditions under which a minimizer of $\hat{Q}_n$ converges in probability to a minimizer of Q.[16] Lemma A.9 in the Appendix shows that the key requirement for such "point-wise" consistency of the set of minimizers is that $\hat{Q}_n$ converges to Q in probability uniformly over compact sets.

5.2 LAM Estimators

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

Theorem 2 (LAM Estimator)**.** *Let Assumptions 2.1, 3.1 – 3.3 and 5.1 – 5.3 hold and the infimum in the risk lower bound be attained within a compact set $K \subset \mathbb{B} \times \mathbb{D}$. Let $\hat{v}_n = (\hat{v}_{1,n}, \hat{v}_{2,n})$ solve*

$$\inf_{(v_1,v_2)\in K} \sup_{s\in\hat{R}_n} \mathbb{E}\left\{ l_M\left(\hat{\phi}_n'(\hat{\mathbb{G}}_n^* + v_1 + s) - \hat{\phi}_n'(s) + v_2\right) \,\middle|\, X_1^n \right\}, \tag{16}$$

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

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

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

is Locally Asymptotically Minimax, that is,

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

Two comments are in order. First, Theorem 2 suggests that the efficient estimator takes a simple form of a "plug-in" estimator with two additive adjustment terms. The role and the numerical values of the optimal adjustment terms depend on the chosen loss function. In particular, for real-valued parameters $\phi(\theta_0)$, choosing the squared loss function allows to select the adjustment terms that balance the bias-variance trade-off. Second, calculating the optimal adjustment terms amounts to solving the optimization problem in (16). This min-max problem may be computationally hard, because the objective function is not convex-concave, and evaluating it at each (v_1, v_2, s) requires bootstrap approximation. However, in some common applications, simple computational heuristics can speed up the optimization, as discussed in Section 5.4.

5.2.1 Euclidean Parameters

Consider $\theta_0 \in \mathbb{R}^{d_\theta}$ and $\phi(\theta) \in \mathbb{R}^{d_\phi}$. Let Σ_θ denote the variance lower bound for θ and $R(\Sigma_\theta)$ denote its range. According to Corollary 1.1 and Remark 1, the risk lower bound is given by

$$\inf_{(v_1,v_2)\in\mathbb{R}^{d_\theta+d_\phi}}\sup_{s\in R(\Sigma_\theta)}\mathbb{E}\left\{\, l_M(\phi_0'(\mathbb{G}_0+v_1+s)-\phi_0'(s)+v_2)\,\right\}. \tag{17}$$

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

$$\begin{aligned} R_n &= \left\{ t = \sum_{j=1}^{d_\theta}\alpha_j\sigma_j \in \mathbb{R}^{d_\theta} \;\middle|\; ||\alpha|| \leqslant \lambda_n \right\}, \\ \hat{R}_n &= \left\{ t = \sum_{j=1}^{d_\theta}\alpha_j\hat{\sigma}_j \in \mathbb{R}^{d_\theta} \;\middle|\; ||\alpha|| \leqslant \lambda_n \right\}. \end{aligned} \tag{18}$$

Then, $\hat{R}_n$ and R_n satisfy Assumption 5.3, and the following Corollary holds.

Corollary 2.1 (LAM Estimation of Euclidean Parameters)**.** *Consider* $\theta_0 \in \mathbb{R}^{d_\theta}$*,* $\phi : \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_\phi}$*. Let Assumptions 2.1, 3.1 - 3.3, 5.1 (i), and 5.2 hold with* $\mathbb{B} = \mathbb{R}^{d_\theta}$ *and* $\mathbb{D} = \mathbb{R}^{d_\phi}$*; define* $\hat{R}_n$ *as in Equation (18). Assume that the infimum in (17) is attained within some compact set* $K \subset \mathbb{R}^{d_\theta + d_\phi}$ *and let* $(\hat{v}_{1,n}, \hat{v}_{2,n})$ *solve*

$$\inf_{(v_1,v_2)\in K} \sup_{s\in\hat{R}_n} \mathbb{E}\left\{ l_M(\hat{\phi}'_n(\mathbb{G}^*_n + v_1 + s) - \hat{\phi}'_n(s) + v_2) \middle| X^n_1 \right\} \tag{19}$$

If Σ_θ *is full-rank, the supremum in (19) can be taken over* $\mathbb{R}^{d_\theta}$*. Then, the estimator sequence*

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

is Locally Asymptotically Minimax. That is,

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

5.2.2 Infinite-Dimensional Parameters

Next, consider estimating the support $S(\mathbb{G}_0)$ according to Assumption 5.3 in the settings when $\theta \in \mathbb{B}$ is infinite-dimensional. I will discuss two different approaches.

The first approach is "brute-force" and uses the fact that $S(\mathbb{G}_0)$ equals the closure of $\theta'_0(T(P))$ in $\mathbb{B}$. Let $g_1, g_2, \ldots$ denote a complete sequence in $L_2(P)$, in a sense that for any $f \in L_2(P)$ and any $\varepsilon > 0$, there exist an $m \in \mathbb{N}$, and $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ such that $||f - \sum_{j=1}^m \alpha_j g_j||_{2,P} < \varepsilon$. For example, the space of continuous functions supported on compact sets is dense in $L_2(P)$, and the space of polynomials is dense within that space, by the Stone-Weierstrass theorem. Therefore, $g_1, g_2, \ldots$ can be chosen as properly truncated polynomials. The idea is to use the g_j-s to construct a sequence of compact sieves in the closure of $\theta'_0(T(P))$. To illustrate, suppose that $T(P) = L^0_2(P)$. Let $h_j = g_j - \mathbb{E}_P(g_j)$ denote the projection of g_j onto $L^0_2(P)$, and $\hat{h}_j = g_j - n^{-1}\sum_{i=1}^n g_j(X_i)$ be its sample analog. Let $\hat{\theta}'_n : L^0_2(P) \to \mathbb{B}$ be a suitable estimator for the path-wise derivative map, and define, for $l_n \in \mathbb{N}$ and $\lambda_n \in \mathbb{R}_+$, the sets

$$\begin{aligned}\hat{R}_n &= \left\{ \sum_{j=1}^{l_n} \alpha_j \hat{\theta}'_n(\hat{h}_j) \;\middle|\; ||\alpha|| \leqslant \lambda_n \right\}, \\ R_n &= \left\{ \sum_{j=1}^{l_n} \alpha_j \theta'_0(h_j) \;\middle|\; ||\alpha|| \leqslant \lambda_n \right\}.\end{aligned} \tag{20}$$

The following Lemma provides primitive conditions under which $\hat{R}_n$ and R_n defined above satisfy Assumption 5.3.

Lemma 1 (Estimating the Support via Projections). *Assume that:*

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

2. $\lambda_n \cdot \sqrt{\frac{l_n}{n}} \cdot \max_{j \leqslant l_n} ||g_j||_{2,P} = o(1)$

Then $\hat{R}_n$ and R_n defined in (20) satisfy Assumption 5.3.

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

The second approach is similar in spirit to the Euclidean case, and is based on characterizing the support of a Gaussian process $\mathbb{G}_0$ via its covariance kernel. The main idea is illustrated below in the example where $\mathbb{G}_0$ is a Gaussian process with $S(\mathbb{G}_0) = C_b([0,1])$ endowed with the sup-norm. The technical details are deferred to Remark 2. Let $K : [0,1] \times [0,1] \to \mathbb{R}$ defined by $K(s,t) = \mathbb{E}(\mathbb{G}_0(t)\mathbb{G}_0(s))$ denote the covariance kernel of $\mathbb{G}_0$, and $\hat{K}_n$ denote a suitable estimator. Denote:

$$\begin{aligned}R_n &= \left\{ f(s) = \sum_{j=1}^{l_n} \alpha_j K(t_j, s) \;\middle|\; 0 \leqslant t_1 < \cdots < t_{l_n} \leqslant 1;\ ||\alpha|| \leqslant \lambda_n \right\}, \\ \hat{R}_n &= \left\{ f(s) = \sum_{j=1}^{l_n} \alpha_j \hat{K}(t_j, s) \;\middle|\; 0 \leqslant t_1 < \cdots < t_{l_n} \leqslant 1;\ ||\alpha|| \leqslant \lambda_n \right\}.\end{aligned} \tag{21}$$

The following Lemma sprovides a primitive condition under which $\hat{R}_n$ and R_n satisfy Assumption 5.3.

Lemma 2 (Estimating the Support via Covariance Kernel)**.** *Let R_n and $\hat{R}_n$ be defined in (21) with $l_n \in \mathbb{N}$, and $\lambda_n \in \mathbb{R}_+$. Suppose that $\hat{K}_n : [0,1] \times [0,1] \to \mathbb{R}$ satisfies*

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

Then, R_n and $\hat{R}_n$ satisfy Assumption 5.3.

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

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

By Theorem 3.2.3 in Bogachev (1998), for each $b^* \in \mathbb{B}_\gamma^*$, there is $h_{b^*} \in H(\gamma)$ such that $K(b^*, c^*) = c^*(h_{b^*})$ for all $c^* \in \mathbb{B}_\gamma^*$. In this sense, every element of $\mathbb{B}_\gamma^*$ can be associated with a unique element of $H(\gamma)$. Therefore, the set $H(\gamma)$ can be mapped out by choosing different b^* and finding the associated h_b^*.

For example, let $\mathbb{B}_0 = C_b([0,1])$ be a set of continuous bounded functions on $[0,1]$, and $\mathbb{G}_0$ denote a Gaussian process with covariance kernel $K(s,t) \equiv \mathbb{E}(\mathbb{G}_0(s)\mathbb{G}_0(t))$. Recall that the continuous dual $\mathbb{B}_0^*$ is the set of all finite Borel measures on $[0,1]$ so that $b^*(x) = \int x(t) d\mu_{b^*}(t)$. With the help of Fubini's theorem, one can verify that $h_{b^*}(s) = \int K(s,t) d\mu_{b^*}(t)$. Further, the set of finitely-supported Borel measures $\{\sum_{j=1}^J \alpha_j \delta_{t_j} : \alpha_j \in \mathbb{R}, t_j \in [0,1], J \in \mathbb{N}\}$, where δ_t denotes the Dirac measure with mass at t, is weak-star dense in $\mathbb{B}_0^*$ meaning that any such $h_{b^*}(s)$ can be approximated by a sequence of the form $\sum_j \alpha_j K(s, t_j)$ point-wise in s, and therefore uniformly since $s \in [0,1]$. This motivates the definition of R_n in (21).

5.3 Examples Revisited

Example 1. Focus on the upper bound $\phi(\theta_0) = \min_{j \leqslant d}(\theta_{0,j})$ with $\theta_0 \in \mathbb{R}^d$. Here, estimating the directional derivative (see Equation 7) amounts to selecting θ_j that are sufficiently close to each other, which is essentially an inequality selection problem. One way to proceed is to test a set of hypotheses $H_0 : \theta_{0,j} \leqslant \theta_{0,i}$ for all i, j (following e.g. Romano, Shaikh, and Wolf (2014)), collect all j-s for which the null is not rejected into the set $\hat{B}_n$, and set

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

Then, if the test size approaches zero as n approaches infinity, $\hat{\phi}'_n(h) = \phi'_0(h)$ for all $h \in \mathbb{R}^d$, with probability approaching one, so that the resulting estimator satisfies Assumption 5.2. ■

Example 3. Suppose again that $N = 2$, and let $\hat{\theta}_n = (\psi_1(\hat{G}_{1:2}), \psi_2(\hat{G}_{2:2}))$ where $\hat{G}_{j:2}$, for $j = 1, 2$ are the empirical CDFs of order statistics of bids. The form of the directional derivative in Equation (23) suggests a natural sample counterpart. For a positive sequence $\kappa_n \downarrow 0$, define the sets

$$\begin{aligned} \hat{S}_{1,n} &= \{v \in [\underline{v}, \overline{v}] : \psi_1(\hat{G}_{1:2}(v)) < \psi_2(\hat{G}_{2:2}(v)) - \kappa_n\}, \\ \hat{S}_{2,n} &= \{v \in [\underline{v}, \overline{v}] : \psi_2(\hat{G}_{2:2}(v)) < \psi_1(\hat{G}_{1:2}(v)) - \kappa_n\}, \\ \hat{S}_{0,n} &= \{v \in [\underline{v}, \overline{v}] : |\psi_1(\hat{G}_{1:2}(v)) - \psi_2(\hat{G}_{2:2}(v))| \leqslant \kappa_n\}, \end{aligned} \tag{22}$$

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

$$\hat{\phi}'_n(h)(v) = h_1(v)\mathbf{1}(v \in \hat{S}_{1,n}) + h_2(v)\mathbf{1}(v \in \hat{S}_{2,n}) + \min(h_1(v), h_2(v))\mathbf{1}(v \in \hat{S}_{0,n}). \tag{23}$$

Then, if $\kappa_n\sqrt{n} \to \infty$, one can show that the resulting estimator satisfies Assumption 5.2 even with R_n^δ replaced by $D([\underline{v}, \overline{v}], [0, 1])^2$. ■

5.4 Computation

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

The main factor that slows down the relevant optimization problem in (19) is that the objective function is costly to evaluate. The approach discussed below aims to reduce the number of evaluations. I focus on the finite-dimensional parameters for simplicity, but similar ideas can be applied in infinite-dimensional settings as well, after selecting suitable sieves. The lower bound from Corollary 1.1 can be equivalently written as

$$\inf_{(v_1,v_2)\in\mathbb{R}^{d_\theta+d_\phi}} \sup_{s\in B} \sup_{\lambda\geqslant 0} \mathbb{E}\left\{l\left(\phi_0'(Z+\lambda s+v_1)-\lambda\phi_0'(s)+v_2\right)\right\}, \tag{24}$$

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

$$g(\lambda)=\mathbb{E}\left\{l\left(\phi_0'(Z+\lambda s+v_1)-\lambda\phi_0'(s)+v_2\right)\right\}$$

that traces the value of the objective function along the ray passing through s. A useful property that appears to hold in practice but turns out to be hard to prove theoretically is that $g(\lambda)$ is maximized at zero or infinity. Therefore, for each (v_1, v_2), the supremum can be calculated by selecting a set of directions (i.e., values of s) on the unit ball and evaluating the function $g(\lambda)$ at zero and some large value of the argument in each direction. Since the directional derivative is typically a partially linear function with a small number of different slopes, this approach allows to reduce the number of evaluations of the objective function dramatically.

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

$$v^* = \frac{1}{2}\max_{I\subset\{1,\ldots,d\}}\left(\frac{\mathbb{E}((\max_{j\in I} Z_i)^2)-\mathbb{E}(Z_{i^*}^2)}{\mathbb{E}(\max_{j\in I} Z_j)}\right), \tag{25}$$

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

$$v^* = \frac{1}{2} \min_{I \subset \{1,\dots,d\}} \left(\frac{\mathbb{E}((\min_{j \in I} Z_i)^2) - \mathbb{E}(Z_{i^*}^2)}{\mathbb{E}(\min_{j \in I} Z_j)} \right). \tag{26}$$

Extensive simulations suggest that these closed-form adjustment terms actually attain the global minimum in (24), although the corresponding formal result is hard to establish. Recalling that $Z \sim N(0, \Sigma)$ with Σ consistently estimated by $\hat{\Sigma}_n$ suggests the following procedure: (i) draw $Z_1^*, \dots, Z_B^* \sim N(0, \hat{\Sigma}_n)$ for some large B and (ii) replace expectations with sample averages in the expressions above.

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

$$\begin{aligned}
&\inf_{(v_1,v_2) \in \mathbb{R}^{d_1+d_2+2}} \sup_{s \in \mathbb{R}^{d_1+d_2}} \mathbb{E}\left\{ l\left(\phi_0'(Z + s + v_1) - \phi_0'(s) + v_2\right)\right\} \\
&\qquad = \inf_{(v_{11},v_{12}) \in \mathbb{R}^{d_1+1}} \sup_{s_1 \in \mathbb{R}^{d_1}} \mathbb{E}\left\{ \left(\min(Z_1 + s_1 + v_{11}) - \min(s_1) + v_{12}\right)^2 \right\} \\
&\qquad\qquad + \inf_{(v_{21},v_{22}) \in \mathbb{R}^{d_2+1}} \sup_{s_2 \in \mathbb{R}^{d_1}} \mathbb{E}\left\{ \left(\max(Z_2 + s_2 + v_{21}) - \max(s_2) + v_{22}\right)^2 \right\}.
\end{aligned}$$

Then, $v_{12}^* = v_{22}^* = 0$ and the approximate solutions (v_{11}^*, v_{21}^*) to each of the problems are given by equations (26) and (25) correspondingly. Similar arguments can be applied in the setting of Example 3 if the loss function $l : D([\underline{v}, \overline{v}]) \to \mathbb{R}_+$ is given by $l(x) = \sum_{i=1}^d x(v_i)^2$ for some fixed $v_1, \dots, v_d \in [\underline{v}, \overline{v}]$.

6 Simulation Study

I illustrate the finite-sample performance of the proposed LAM estimator by comparing it with the simple "plug-in" estimator and the existing bias correction approaches. For simplicity, I focus on the upper bound from Example 1: $\phi(\theta) = \min_{j \leqslant d}(\theta_j)$ with

$\theta \in \mathbb{R}^d$. The results for the lower bound and for both bounds together are similar.

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

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

Kreider and Pepper (2007) found that this method performs well in practice, even though it is not theoretically supported.[17] Studying the asymptotic properties of such estimator is beyond the scope of this paper.

The second approach is due to Chernozhukov, Lee, and Rosen (2013). The authors propose a half-median unbiased estimator which lies above the true value with probability at least one half asymptotically.[18] The estimator takes the form

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

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

Next, consider the implementation of the LAM estimator. Let $Z \sim N(0, \Sigma)$, denote the weak limit of the efficient estimator sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$. To approximate the law of Z in accord with Assumption 5.1, one may pick a consistent estimator $\hat{\Sigma}_n$ for Σ and chose Z_n^* to be a random vector distributed as $N(0, \hat{\Sigma}_n)$, conditional on the data. To construct a suitable estimator for the directional derivative, one may

[17] The standard arguments for consistency of the procedure rely on the differentiability of the function ϕ, which, in the present setting, may fail. See Tibshirani and Efron (1993).

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

follow the procedure described in Section 5.3 and obtain $\hat{\phi}'_n(h) = \min_{j \in \hat{B}_n}(h_j)$. Then, calculate the adjustment term $\hat{v}_{1,n}$ by minimizing

$$\inf_{v_1 \in \mathbb{R}^d} \sup_{c \in \mathbb{R}^d} \mathbb{E}\left((\hat{\phi}'_n(Z^*_n + v_1 + c) - \hat{\phi}'_n(c))^2 \middle| X^n_1 \right)$$

and set

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

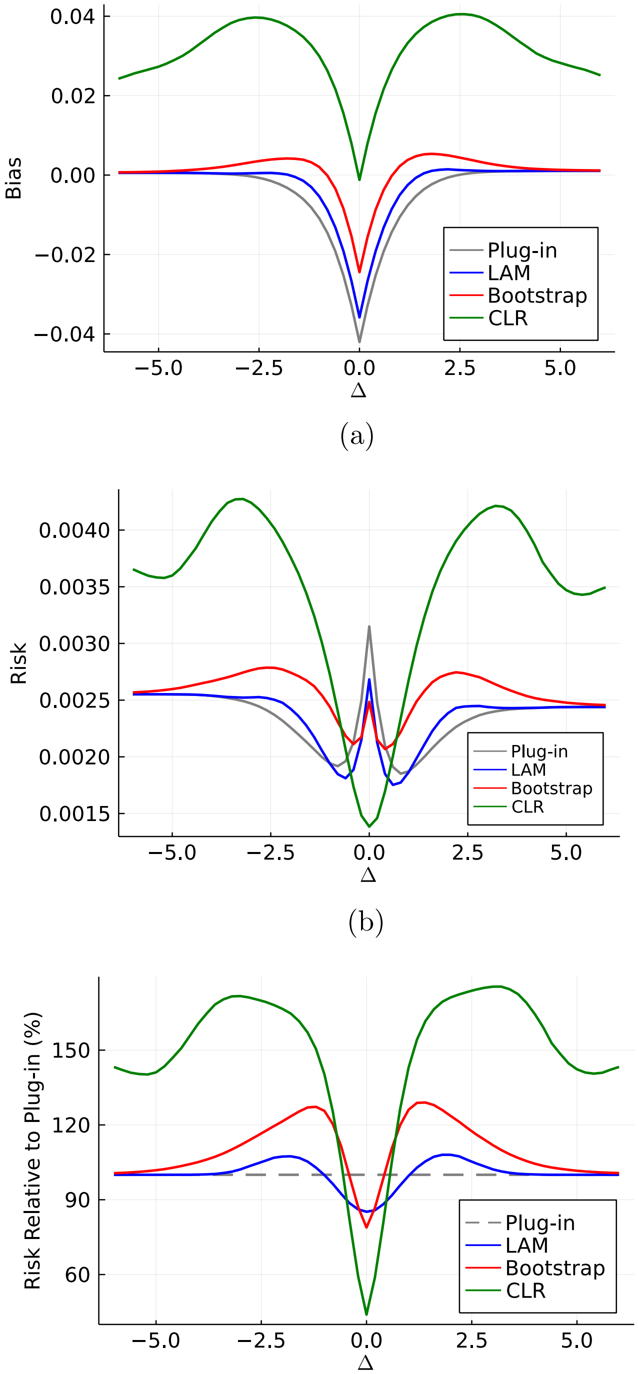
In this special case the second adjustment term is not required and the optimization problem is simplified. Moreover, the squared loss function allows to choose $\hat{v}_{1,n}$ to balance the bias-variance trade-off, and compute approximate closed-form solutions as discussed in Section 5.4.

The simulation setup is as follows. The data $X_1, \ldots, X_n$ are i.i.d. from $N(\theta_0, \Sigma)$ in $\mathbb{R}^3$, so that $\theta_0 = \mathbb{E}_P(X)$. I consider an ordinary covariance matrix Σ with different variances and non-zero correlations, and set $\theta(\Delta) = (0, \Delta/\sqrt{n}, 2\Delta/\sqrt{n})^T$ so that Δ plays the role of the local parameter. That is, Δ equal to zero corresponds to the point $\theta_0 = (0,0,0)^T$, where the full differentiability of ϕ fails, and varying Δ allows to "walk across" the local neighborhood of this point.[19] For each value of the local parameter Δ on a grid chosen to scale, I perform $M = 5000$ simulations, with $B = 2000$ bootstrap draws and sample size $n = 300$. For every draw, indexed by m, I generate a random sample $X^m_1, \ldots, X^m_n$ from $N(\theta_0, \Sigma)$, and calculate $\hat{\phi}^{\text{Plug-in}}_m = \phi(\bar{X}_m)$, and $\hat{\phi}^{\text{Bootstrap}}_m$, $\hat{\phi}^{\text{CLR}}_m$ and $\hat{\phi}^{\text{LAM}}_m$ according to the formulas above. Then, I compute the average bias, $\frac{1}{M}\sum_{m=1}^{M}(\hat{\phi}_m - \phi(\theta(\Delta)))$, and risk, $\frac{1}{M}\sum_{m=1}^{M}(\hat{\phi}_m - \phi(\theta(\Delta)))^2$, for each of the four estimators and plot the results as a function of Δ.

The results presented in Figure 2 require several comments. First, Panel (a) suggests that the LAM estimator does not reduce the bias as much as the other methods. This is not surprising, since the LAM estimator was constructed targeting the mean-squared error (i.e., variance plus bias squared), rather than the bias directly. Larger reduction in bias can be achieved by using a different loss function, such as $l(x) = |x|^\alpha$ for $0 < \alpha < 2$. Second, Panel (b) suggests that the LAM estimator has the lowest worst-case risk, which is consistent with the asymptotic results of Theorems 1 and 2. Note that while the risk of the plug-in estimator is maximized

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

Figure 2: Finite-Sample Bias, Risk, and Relative Risk.



(a)

(b)

(c)

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

at zero (i.e., at the point of non-differentiability) the maximum risks of the bias-corrected estimators are attained away from zero. Moreover, the LAM estimator outperforms the bias-correction methods in terms of risk everywhere except for a small neighborhood of zero. Finally, Panel (c) shows relative risks in percentage terms, suggesting that the bias-corrected estimators may have a substantially larger risk than the Plug-in, depending on the value of Δ, while the LAM estimator does not. Since Δ is unknown and cannot be consistently estimated, the LAM estimator can be interpreted as cautious.

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

7 English Auctions with IPV

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

7.1 Model and Identification

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

7.1.1 CDF of Valuations

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

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

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

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

It is well-known that the distribution of any order statistic of a collection of i.i.d. random variables uniquely determines the parent distribution: for each $j \leqslant N$, there is a strictly increasing and differentiable function $\psi_j : [0,1] \to [0,1]$ such that $F(v) = \psi_j(F_{j:N}(v))$[20]. Applying ψ_j to both sides of the two previous displays for every $j \leqslant N$ and intersecting the results, Haile and Tamer (2003) obtain the following point-wise bounds:

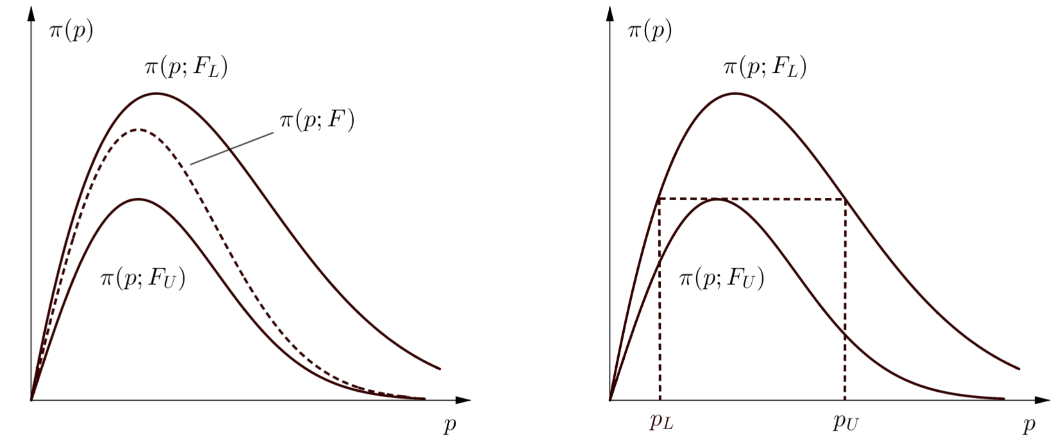
$$\psi_{N-1}(G_{N:N}(v - \Delta)) \leqslant F(v) \leqslant \min_{j \leqslant N} \psi_j(G_{j:N}(v)). \tag{27}$$

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

7.1.2 Optimal Reserve Price

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

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



(a) Bounds on the true profit function (b) Implied bounds on the maximizer

Figure 3: Identification of the Optimal Reserve Price.

price maximizes

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

where v_0 denotes the value of the unsold good to the seller. Denoting the bounds on the CDF by $F_L(v) \leqslant F(v) \leqslant F_U(v)$, it follows that $\pi(p; F_U) \leqslant \pi(p; F) \leqslant \pi(p; F_L)$ for all p. As illustrated in Figure 3, this implies the following bounds $[p_L, p_U]$ on the optimal reserve price:

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

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

7.2 Estimation

It is assumed that the researcher observes an i.i.d. sample of auction data which includes bids $\{B_i\}_{i=1}^n$ where $B_i = (B_{1,i}, \ldots, B_{N,i})$. Such data can be used to estimate the empirical CDFs of order statistics of bids.[21] Consider estimating the upper bound on the distribution of valuations from Equation (27). For a fixed $v \in [\underline{v}, \overline{v}]$, the upper bound takes the form $\phi(\theta(v)) = \min_{j \leqslant d}(\theta_d(v))$, where $\theta(v)$ is a vector of smooth transformations of the CDFs of bids evaluated at v. Haile and Tamer (2003) propose to approximate the minimum by a sequence of smooth functions, chosen to reduce the finite-sample bias. Specifically, they consider the function

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

where ρ is the smoothness parameter. This function satisfies $\tilde{\phi}(\theta; \rho) > \min_{j \leqslant d}(\theta_j)$ for any $\rho \in \mathbb{R}$, and $\lim_{\rho \to -\infty} \mu(\theta; \rho) = \min_{j \leqslant d}(\theta_j)$. Letting $\hat{\theta}_n$ denote an estimator for θ_0 and $\rho_n \to -\infty$ denote an appropriate sequence of smoothing parameters,[22] they set

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

Such estimator has the same asymptotic properties as $\hat{\phi}_n^{\text{Plug-in}} = \min_{j \leqslant d}(\hat{\theta}_{j,n})$, with the advantage of providing bias-correction in finite-samples.[23]

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

[21] The analysis can be performed conditional on auction characteristics and the number of participants. To apply the results of this paper, the auction characteristics must be discrete (or discretized) to ensure that the conditional CDF-s of the bids can be regularly estimated (see Section 3.2.2). Note that, since the IPV assumption is imposed conditional on the auction characteristics, focusing on discrete characteristics may be restrictive.

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

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

as discussed below.

The construction of LAM estimator in this setting has been discussed in Example 3 throughout the paper. The parameter of interest is a pair of CDF-type functions, $\phi(\theta_0) \in D([\underline{v}, \overline{v}], [0,1])^2$, representing the bounds on F in Equation (27). To focus on the bias-variance trade-off in estimation and simplify the computation of the adjustment terms, I consider the squared loss function that focuses on a finite grid of points $v_1, \dots, v_K \in [\underline{v}, \overline{v}]$. Specifically, the loss function $l : D([\underline{v}, \overline{v}], [0,1])^2 \to \mathbb{R}_+$ is given by $l(x_1, x_2) = \sum_{k=1}^{K} (x_1(v_k)^2 + x_2(v_k)^2)$. Then, as discussed in Section 5.4, the optimization problem can be split into several simple subproblems that have approximate closed-form solutions.

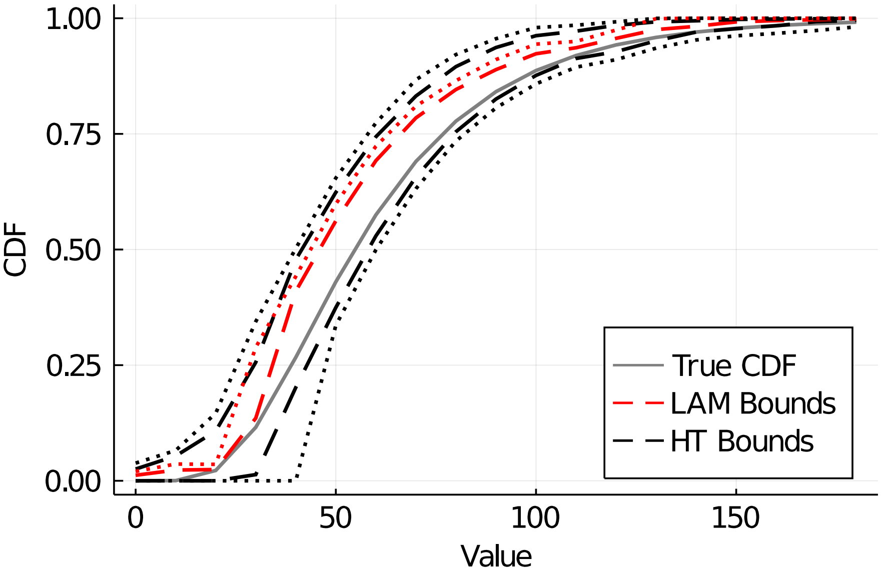
7.3 Results

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

Figure 4 presents the results. First, since the lower bound equals to $\psi_{N-1}(G_{N:N}(v - \Delta))$, no smoothing or adjustment is required and the two estimation methods yield the same results. The estimated lower bound is fairly tight throughout the support since the minimal bid increment is relatively small and jump bidding is not too common. Second, the LAM estimates for the upper bound are, on average, substantially tighter than the HT estimates. In particular, the 95th quantile for the LAM estimate (red dotted line) is consistently below the average HT estimate (black dashed line) across simulations. At the same time, there is some downward bias in the LAM estimates around the lower part of the support. This issue, caused by the fact that the highest bids in that region are very rarely observed, disappears with smaller N and/or sufficiently large n.

Table 1 presents the implied bounds on the optimal reserve prices for different parameters of the Log-Normal distribution. While the bounds estimated with both methods are fairly wide, the LAM estimates are, on average, substantially tighter.

Figure 4: Estimated Bounds on the CDF of Valuations



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

Table 1: Estimated Bounds on the Optimal Reserve Price

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

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

8 Extension to Convex Cones

In the settings where the tangent set $T(P)$ is a convex cone, the lower bound in Theorem 1 holds with $S(\mathbb{G}_0)$ replaced by $\theta_0'(T(P))$. Such settings typically arise in the presence of moment inequality restrictions that are binding at P. Common examples include point- or over-identifying moment inequality models, or regression models with binding sign constraints. However, such settings are theoretically problematic: when $T(P)$ is a convex cone, the optimal estimators proposed by Convolution and Minimax Theorems may often be inadmissible, even for differentiable parameters.[24] To illustrate, I consider a simple example, similar to Imbens and Manski (2004).

Suppose that the parameter of interest $\theta_0 \in \mathbb{R}$ is partially identified, and the bounds are given by $\theta_{L,0} = \theta_L(P)$ and $\theta_{U,0} = \theta_U(P)$, which are "smooth" functionals (i.e., differentiable in the sense of Definition 3.3) of the distribution P of the observable random vector X. The model is given by:[25]

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

What is an efficient estimator for the identified set $[\theta_{L,0}, \theta_{U,0}]$? In this example, stimating the identified set amounts to estimating a two-dimensional vector of bounds. First, consider a situation when $\theta_L(P) < \theta_U(P)$. In this case, the tangent set is unrestricted, i.e., $T(P) = L_2^0(P)$, and the classical efficiency theory suggests that the "plug-in" estimator, defined by $\hat{\theta}_{L,n} \equiv \theta_L(\hat{P}_n)$ and $\hat{\theta}_{U,n} \equiv \theta_U(\hat{P}_n)$, where $\hat{P}_n$ denotes the empirical distribution, is optimal. Intuitively, the bounds can be estimated separately because they are not informative about each other. On the other hand, suppose that $\theta_L(P) = \theta_U(P)$. In this case, the estimators $\hat{\theta}_{L,n}$ and $\hat{\theta}_{U,n}$ target the same parameter, so the intuition suggests that they may be combined to produce a more efficient estimator. For example, assuming that the asymptotic variances of $\hat{\theta}_{L,n}$ and $\hat{\theta}_{U,n}$ are the same, the optimal GMM would suggest using $(\hat{\theta}_{L,n} + \hat{\theta}_{U,n})/2$ to estimate both θ_L and θ_U. However, due to the tangent set being a cone, the existing semiparametric efficiency theory suggests otherwise. More precisely, denoting the path-wise derivatives by $\theta_{0,L}'(h) = \mathbb{E}_P(\psi_L h)$ and $\theta_{0,U}'(h) = \mathbb{E}_P(\psi_U h)$ for some

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

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

$\psi_L, \psi_U \in L_2^0(P)$, the tangent set is given by

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

Then, since $\overline{\text{lin}}\ T(P) = L_2^0(P)$, both the Convolution Theorem (Theorem A.5) and LAM Theorem (Theorem 1, applied with $\phi(\theta_L, \theta_U) = (\theta_L, \theta_U)$) suggest that the "plug-in" estimator $[\hat{\theta}_{L,n}, \hat{\theta}_{U,n}]$ is still optimal, which contradicts the above intuition.[26]

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

9 Conclusion

In many econometric models, certain parameters of interest are represented via directionally differentiable functionals. The potential lack of full differentiability has raised concerns in regard to choosing "good" estimators for such parameters. This paper proposed a solution by extending the classical Local Asymptotic Minimax Theorem to a class of directionally differentiable parameters. First, I derived the general risk lower bound that covers all reasonable estimators and holds for a variety of loss functions. In contrast to the fully differentiable settings, the optimal estimator depends on the chosen loss function, suggesting that it must be tailored to specific applications. Second, I showed that the optimal estimator takes a relatively simple form of the "plug-in" estimator with additive adjustment terms and provided a general procedure to compute them from the data. It typically does not reduce the bias as much as some of the existing methods, but helps to avoid large fluctuations in risk around the points where the full differentiability fails. The empirical relevance of the proposed method was demonstrated in the application to English auctions with independent private values.

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

References

Ackerberg, D., X. Chen, J. Hahn, and Z. Liao (2014): "Asymptotic efficiency of semiparametric two-step GMM," *Review of Economic Studies*, 81(3), 919–943.

Ai, C., and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

——— (2012): "The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions," *Journal of Econometrics*, 170(2), 442–457.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile regression under misspecification, with an application to the US wage structure," *Econometrica*, 74(2), 539–563.

Aradillas-López, A., A. Gandhi, and D. Quint (2013): "Identification and inference in ascending auctions with correlated private values," *Econometrica*, 81(2), 489–534.

Arnold, B. C., N. Balakrishnan, and H. N. Nagaraja (2008): *A first course in order statistics*. SIAM.

Artstein, Z. (1983): "Distributions of random sets and random selections," *Israel Journal of Mathematics*, 46(4), 313–324.

Bassett, G., and R. Koenker (1982): "An empirical quantile function for linear models with iid errors," *Journal of the American Statistical Association*, 77(378), 407–415.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79(6), 1785–1821.

Beresteanu, A., and F. Molinari (2008): "Asymptotic properties for a class of partially identified models," *Econometrica*, 76(4), 763–814.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and adaptive estimation for semiparametric models*, vol. 4. Johns Hopkins University Press Baltimore.

Blundell, R., A. Gosling, H. Ichimura, and C. Meghir (2007): "Changes in the distribution of male and female wages accounting for employment composition using bounds," *Econometrica*, 75(2), 323–363.

Bogachev, V. I. (1998): *Gaussian measures*, no. 62. American Mathematical Soc.

Bogachev, V. I. (2007): *Measure theory*, vol. 1. Springer Science & Business Media.

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set identified linear models," *Econometrica*, 80(3), 1129–1155.

Brown, B. W., and W. K. Newey (1998): "Efficient semiparametric estimation of expectations," *Econometrica*, 66(2), 453–464.

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of econometrics*, 34(3), 305–334.

——— (1992): "Efficiency bounds for semiparametric regression," *Econometrica: Journal of the Econometric Society*, pp. 567–596.

Chen, X., and A. Santos (2018): "Overidentification in regular models," *Econometrica*, 86(5), 1771–1817.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): "Quantile and probability curves without crossing," *Econometrica*, 78(3), 1093–1125.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: estimation and inference," *Econometrica*, 81(2), 667–737.

Chesher, A., and A. M. Rosen (2017): "Generalized instrumental variable models," *Econometrica*, 85(3), 959–989.

Ciliberto, F., and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77(6), 1791–1828.

Conway, J. B. (1985): *A Course in Functional Analysis*. Springer, New York, NY.

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

Dvoretzky, A., A. Wald, and J. Wolfowitz (1951): "Elimination of randomization in certain statistical decision procedures and zero-sum two-person games," *The Annals of Mathematical Statistics*, pp. 1–21.

Fang, Z. (2018): "Optimal plug-in estimators of directionally differentiable functionals," *Working paper*.

Fang, Z., and A. Santos (2019): "Inference on directionally differentiable functions," *The Review of Economic Studies*, 86(1), 377–412.

Feinberg, E. A., and A. B. Piunovskiy (2006): "On the Dvoretzky–Wald–Wolfowitz theorem on nonrandomized statistical decisions," *Theory of Probability & Its Applications*, 50(3), 463–466.

Galichon, A., and M. Henry (2011): "Set Identification in Models with Multiple Equilibria," *The Review of Economic Studies*, 78(4), 1264–1298.

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

Hansen, B. E. (2017): "Regression kink with an unknown threshold," *Journal of Business & Economic Statistics*, 35(2), 228–240.

Hirano, K., and J. R. Porter (2009): "Asymptotics for statistical treatment rules," *Econometrica*, 77(5), 1683–1701.

——— (2012): "Impossibility results for nondifferentiable functionals," *Econometrica*, 80(4), 1769–1790.

Ho, K., and A. M. Rosen (2015): "Partial identification in applied research: benefits and challenges," Discussion paper, National Bureau of Economic Research.

Hong, H., and J. Li (2020): "The numerical bootstrap," *The Annals of Statistics*, 48(1), 397–412.

Horowitz, J. L. (2001): "The bootstrap," in *Handbook of econometrics*, vol. 5, pp. 3159–3228. Elsevier.

Ibragimov, I. A., and R. Z. Hasḿinskii (1981): *Statistical estimation: asymptotic theory*. Springer, New York, NY.

Imbens, G. W., and C. F. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72(6), 1845–1857.

Kaido, H., and A. Santos (2014): "Asymptotically efficient estimation of models defined by convex moment inequalities," *Econometrica*, 82(1), 387–413.

Kline, P., and A. Santos (2013): "Sensitivity to Missing Data Assumptions: Theory and Evaluation of the US Wage Structure," *Quantitative Economics*, 4(2), 231–267.

Koshevnik, Y. A., and B. Y. Levit (1976): "On a non-parametric analogue of the information matrix," *Teoriya Veroyatnostei i ee Primeneniya*, 21(4), 759–774.

Kreider, B., and J. V. Pepper (2007): "Disability and employment: Reevaluating the evidence in light of reporting errors," *Journal of the American Statistical Association*, 102(478), 432–441.

Kreider, B., J. V. Pepper, C. Gundersen, and D. Jolliffe (2012): "Identifying the effects of SNAP (food stamps) on child health outcomes when participation is endogenous and misreported," *Journal of the American Statistical Association*, 107(499), 958–975.

Le Cam, L. M. (1986): *Asymptotic methods in statistical decision theory*. Springer Verlag.

Lehmann, E. L., and G. Casella (2006): *Theory of point estimation*. Springer Science & Business Media.

Manski, C. F., and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997–1010.

——— (2009): "More on monotone instrumental variables," *The Econometrics Journal*, 12, S200–S216.

Manski, C. F., and E. Tamer (2002): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70(2), 519–546.

Masten, M. A., and A. Poirier (2020): “Inference on breakdown frontiers,” *Quantitative Economics*, 11(1), 41–111.

Myerson, R. B. (1981): “Optimal auction design,” *Mathematics of operations research*, 6(1), 58–73.

Newey, W. K. (1990): “Semiparametric efficiency bounds,” *Journal of applied econometrics*, 5(2), 99–135.

Newey, W. K. (1994): “The asymptotic variance of semiparametric estimators,” *Econometrica: Journal of the Econometric Society*, pp. 1349–1382.

Pakes, A. (2010): “Alternative models for moment inequalities,” *Econometrica*, 78(6), 1783–1822.

Pakes, A., J. Porter, K. Ho, and J. Ishii (2007): “Moment inequalities and their application,” Discussion paper, CEMMAP Working paper.

——— (2015): “Moment inequalities and their application,” *Econometrica*, 83(1), 315–334.

Pfanzagl, J. (1982): *Contributions to a General Asymptotic Statistical Theory*. Springer, New York, NY.

Romano, J. P., A. M. Shaikh, and M. Wolf (2014): “A practical two-step method for testing moment inequalities,” *Econometrica*, 82(5), 1979–2002.

Shaikh, A. M., and E. J. Vytlacil (2011): “Partial identification in triangular systems of equations with binary dependent variables,” *Econometrica*, 79(3), 949–955.

Shapiro, A. (1990): “On concepts of directional differentiability,” *Journal of optimization theory and applications*, 66(3), 477–487.

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

Tibshirani, R. J., and B. Efron (1993): “An introduction to the bootstrap,” *Monographs on statistics and applied probability*, 57, 1–436.

van der Vaart, A. W. (1988): “Statistical estimation in large parameter spaces,” *CWI Tracts*.

——— (1991): “On differentiable functionals,” *The Annals of Statistics*, pp. 178–204.

——— (2000): *Asymptotic statistics*, vol. 3. Cambridge university press.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak convergence.* Springer.

A Appendix: Proofs from the Main Text

A.1 Known Results for Reference

Below, I state several known results in the exact form they are used in the proofs for easier reference.

Theorem A.1 (Riesz Representation). *Let M denote a linear subspace of a Hilbert space $(H, \langle \cdot, \cdot \rangle)$ and $L : M \to \mathbb{R}$ denote a continuous linear functional. Then there is an element $\tilde{l} \in \bar{M}$ such that $L(h) = \langle \tilde{l}, h \rangle$ for all $h \in \bar{M}$.*

Proof. L can be extended to a continuous linear functional on $\bar{M}$ by the Hahn Banach Theorem (Theorem 6.2 in Conway, 1985). The result follows from applying the usual Riesz Representation Theorem (Proposition 3.4 in Conway, 1985). ■

Theorem A.2 (General Le Cam's Third Lemma. Theorem 3.10.7. in van der Vaart and Wellner (1996)). *Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$ and let $X_n : \Omega_n \to \mathbb{D}$ be maps with values in a metric space. Assume that Q_n is contiguous with respect to P_n, and*

$$\left(X_n, \frac{dQ_n}{dP_n} \right) \quad \overset{P_n}{\rightsquigarrow} \quad (X, V).$$

Then $L(B) = \mathbb{E}(1_B(X) \cdot V)$ defines a probability measure and $X_n \rightsquigarrow L$ along Q_n. If X is tight or separable then so is L.

Theorem A.3 (Le Cam's Third Lemma. Example 3.10.8 in van der Vaart and Wellner (1996)). *If*

$$\left(X_n, \log \frac{dQ_n}{dP_n} \right) \quad \overset{P_n}{\rightsquigarrow} \quad N\left(\begin{bmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{bmatrix}, \begin{bmatrix} \Sigma & \tau \\ \tau^T & \sigma^2 \end{bmatrix} \right),$$

then

$$X_n \overset{Q_n}{\rightsquigarrow} N(\mu + \tau, \Sigma).$$

The following results refer to Definitions 3.3 and 3.4.

Theorem A.4 (Convolution Theorem for Euclidean Parameters. Theorem 25.20 in van der Vaart (2000))**.** *Assume that $\theta(P) \in \mathbb{R}^{d_\theta}$ is differentiable relative to a tangent set $T(P)$ with the path-wise derivative $\theta_0' : \bar{T}(P) \to \mathbb{R}^d$. Then, for any regular estimator sequence $\hat{\theta}_n$,*

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

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

To state the Convolution Theorem for infinite-dimensional parameters, some new notation is required. For each $b^* \in \mathbb{B}^*$ (the continuous dual of $\mathbb{B}$), $b^* \circ \theta_0'$ is a continuous linear map from $\bar{T}(P)$ into $\mathbb{R}$. By the Riesz Representation Theorem (Theorem A.1), there is an element $\tilde{\theta}_{b^*} \in \bar{T}(P)$ such that $b^* \circ \theta_0'(h) = \mathbb{E}_P(\tilde{\theta}_{b^*}h)$ for any $h \in \bar{T}(P)$. Such $\tilde{\theta}_{b^*}$ is called the *canonical gradient* of θ in the direction b^*.

Theorem A.5 (Convolution Theorem. Theorem 3.11.2. in van der Vaart and Wellner (1996))**.** *Assume that $\theta(P) \in \mathbb{B}$ is differentiable relative to a tangent set $T(P)$ with the path-wise derivative $\theta_0' : \bar{T}(P) \to \mathbb{B}$. Then, for any regular estimator sequence $\hat{\theta}_n$,*

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

where $\mathbb{G}_0$ is a tight centered Gaussian random element in $\mathbb{B}$ and $\mathbb{W}$ is a tight random element in $\mathbb{B}$ independent from $\mathbb{G}_0$. The distribution of $\mathbb{G}_0$ is such that $(b_1^*, \ldots, b_K^*) \circ \mathbb{G}_0$ is a centered Gaussian random vector with $\mathbb{C}ov(b_i^*(\mathbb{G}_0), b_j^*(\mathbb{G}_0)) = \mathbb{E}(\tilde{\theta}_{b_i^*}\tilde{\theta}_{b_j^*})$ for any $b_1^* \ldots b_K^* \in \mathbb{B}^*$. Moreover, the distribution of $\mathbb{G}_0$ concentrates on the closure of $\theta_0'(T(P))$.

Theorem A.6 (Continuous Mapping Theorem. Theorem 1.3.6. in van der Vaart and Wellner (1996))**.** *Let a map between two metric spaces $g : \mathbb{B} \to \mathbb{D}$ be continuous at every point of a set $\mathbb{B}_0 \subset \mathbb{B}$. If $X_n \rightsquigarrow X$ and X takes its values in $\mathbb{B}_0$, then $g(X_n) \rightsquigarrow g(X)$.*

Theorem A.7 (Prohorov's Theorem. Theorem 1.3.9. in van der Vaart and Wellner (1996))**.** *If the sequence X_n is asymptotically tight and asymptotically measurable, then for any subsequence $X_{n'}$ there is a further sibsequence $X_{n''}$ that converges weakly to a tight Borel law.*

The following result, due to Feinberg and Piunovskiy (2006) plays a crucial role in the proof of Theorem 1 below. To state the result, some new concepts are required. A separable metric space S endowed with its Borel sigma field $\mathcal{S}$ is known as a Borel space. Let $(X, \mathcal{X})$ be a Borel space of *states*. Let $(A, \mathcal{A})$ be a Borel space of *actions*, and $\mathcal{P}(A)$ denote the set of all probability distributions on it. Let $A(x) \in \mathcal{A}$ denote the set of feasible actions for each state $x \in X$. A randomized decision rule π is a mapping that for each state $x \in X$ returns a probability distribution on $(A, \mathcal{A})$, supported on $A(x)$. That is, $\pi : X \to \mathcal{P}(A)$ satisfies $\pi(A(x); x) = 1$ where $\pi(S; x)$ denotes the probability that $\pi(x)$ assigns to a set $S \in \mathcal{A}$. A non-randomized decision rule is defined by a measurable map $\varphi : X \to A$ so that $\pi(\{\varphi(x)\}; x) = 1$. Next, for $j = 1, \ldots, J$, let $\rho_j : X \times A \to \bar{\mathbb{R}}$ denote loss functions, which are assumed to be measurable with respect to the product sigma-field $\mathcal{X} \times \mathcal{A}$ (no other restriction is placed on ρ_j). Let $\mu_1 \ldots \mu_K$ denote probability measures on $(X, \mathcal{X})$. The *risk*, associated with a decision rule π, given a loss function ρ_j and a distribution of states μ_k is defined as

$$R(\pi; \rho_j, \mu_k) = \int\limits_X \int\limits_A \rho_j(x, a) d\pi(a; x) d\mu_k(x)$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

Theorem A.8 (Purification Theorem. Theorem 1 in Feinberg and Piunovskiy (2006))**.** *Let $(X, \mathcal{X})$ denote a Borel space of states, and $\mu_1, \ldots, \mu_K$ denote probability distributions on it. Let $(A, \mathcal{A})$ denote a Borel space of actions. If $\mu_1, \ldots, \mu_K$ are non-atomic, then for any randomized decision rule $\pi : X \to \mathcal{P}(A)$ there is an equivalent nonrandomized decision rule $\varphi : X \to A$. That is,*

$$\int\limits_X \int\limits_A \rho_j(x, a) d\pi(a; x) d\mu_k(x) = \int\limits_X \rho_j(x, \varphi(x)) d\mu_k(x)$$

for $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

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

Lemma A.1 (Suprema of Lower Semi-Continuous Functions In Polish Spaces)**.** *Let (X, ρ) be a separable metric space, $B \subset X$ be an arbitrary non-empty subset and $f : X \to \mathbb{R}$ be a lower semi-continuous function. Then B is separable and*

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

where B° denotes a countable dense subset of B.

Proof. First, I show that B is separable. Let $E = \{e_1, e_2, \dots\}$ denote a countable dense subset of X. Fix $\varepsilon > 0$. Define $E' = \{e_j \in E | \rho(e_j, B) \leqslant \varepsilon/3\} = \{e'_1, e'_2, \dots\}$ which is non-empty since E is dense in X. For every such $e'_j \in E'$ there is $x_j \in B$ with $\rho(e'_j, x_j) \leqslant \rho(e'_j, B) + \varepsilon/3 \leqslant 2\varepsilon/3$. Let B° denote a set of all $x_j \in B$ obtained this way. Since E is dense in X, for any $x \in B$ there is $e_k \in E$ with $\rho(e_k, x) \leqslant \varepsilon/3$. Since $\rho(e_k, B) \leqslant \rho(e_k, x)$ by definition, it must be that $e_k = e'_j$ for some $e'_j \in E'$ and $\rho(e'_j, x) \leqslant \varepsilon/3$. For such e'_j there is $x_j \in B^\circ$ with $\rho(e'_j, x_j) \leqslant 2\varepsilon/3$. By triangle inequality, $\rho(x, x_j) \leqslant \rho(x, e'_j) + \rho(e'_j, x_j) \leqslant \varepsilon$ so that B° is a countable dense subset of B.

For the second part of the statement, it is clear that $\sup_{B^\circ} f(x) \leqslant \sup_B f(x)$. For the reversed inequality, it suffices to show $\sup_B f(x) \leqslant \sup_{B^\circ} f(x) + \varepsilon$ for an arbitrary $\varepsilon > 0$. Pick $x' \in B$ such that $\sup_B f(x) \leqslant f(x') + \varepsilon$. Since B° is dense in B, there is a sequence $(x_n)_{n \geqslant 1} \in B^\circ$ such that $\rho(x_n, x') \to 0$. It follows from lower semi-continuity of f that $\liminf_{n\to\infty} f(x_n) \geqslant f(x')$. Therefore, $\sup_B f(x) \leqslant \liminf_{n\to\infty} f(x_n) + \varepsilon \leqslant \sup_{B^\circ} f(x) + \varepsilon$, and the proof is complete.

■

Lemma A.2 (Uniform Convergence of Lipchitz Functions)**.** *Let (X, ρ) denote a compact metric space and $f_n : X \to \mathbb{R}$ be a uniformly Lipchitz sequence of functions, that is, for some constant C independent of n,*

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

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

Proof. First, I show that f satisfies:

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

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

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

Since δ was arbitrary, the desired conclusion follows.

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

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

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

A.2 Proofs from the Main Text and Auxiliary Lemmas

For any probability measures Q and S let dQ/dS denote the density of the part of Q that is absolutely continuous with respect to S. It is understood that $\log 0 = -\infty$, so the extended logarithm $x \mapsto \log x$ is a continuous bijection of $[0, +\infty)$ into $[-\infty, \infty)$ with a continuous inverse. In particular, the log-likelihood ratio converges weakly on $[-\infty, +\infty)$ if an only if the likelihood ratio converges weakly on $[0, +\infty)$. $\bar{\mathbb{R}}$ denotes the extended real line.

Lemma A.3 (Shifted Likelihood Ratio)**.** *Assume that for each $n \geqslant 1$ $(P_{n,h} : h \in H)$ is a set of probability measures indexed by the elements of (a subset of a) Hilbert space H such that*

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

where, for any $h_1, \dots, h_m \in H$, $(\Delta_{n,h_1}, \dots, \Delta_{n,h_m}) \overset{P_{n,0}}{\rightsquigarrow} (\Delta_{h_1}, \dots, \Delta_{h_m})$ and the latter is a centered Gaussian vector with $\mathbb{C}ov(\Delta_{h_i}, \Delta_{h_j}) = \langle h_i, h_j\rangle$ for $i, j = 1, \dots, m$. Then, the following holds.

1. *For any $h, h' \in H$,*

$$\log \frac{dP_{n,h'+h}}{dP_{n,h'}} \quad \overset{P_{n,h'}}{\rightsquigarrow} \quad \Delta_h - \frac{1}{2} ||h||^2 .$$

2. *In particular, let $h_1, \dots, h_m$ be linearly independent and denote $h(a) = \sum_{j=1}^m a_j h_j$ for some $a = (a_1, \dots, a_m) \in \mathbb{R}^m$. Let Σ be a $m \times m$ matrix with elements $\Sigma_{ij} = \langle h_i, h_j \rangle$. Then, for any $h' \in H$ and any $a \in \mathbb{R}^m$:*

$$\log \frac{dP_{n,h'+h(a)}}{dP_{n,h'}} \quad \overset{P_{n,h'}}{\rightsquigarrow} \quad a^T \Delta - \frac{1}{2} a^T \Sigma a$$

where $\Delta = (\Delta_{h_1}, \dots, \Delta_{h_m})$.

Proof. Fix $h, h' \in H$. Let μ_n be a positive sigma-finite measure that dominates $P_{n,0} + P_{n,h'} + P_{n,h+h'}$ and write $p_{n,\tilde{h}} = dP_{n,\tilde{h}}/d\mu_n$ for $\tilde{h} \in \{0, h', h' + h\}$. Pairwise likelihood ratios (i.e., $p_{n,h}/p_{n,0}$, etc.) are left unspecified when the denominator is zero. By assumption,

$$\begin{bmatrix} \log \frac{p_{n,h'+h}}{p_{n,0}} \\ \log \frac{p_{n,h'}}{p_{n,0}} \end{bmatrix} \quad \overset{P_{n,0}}{\rightsquigarrow} \quad \begin{bmatrix} \Delta_{h'+h} - \frac{1}{2} ||h' + h||^2 \\ \Delta_{h'} - \frac{1}{2} ||h'||^2 \end{bmatrix} .$$

Note that the limit concentrates on $\mathbb{R} \times \mathbb{R}$. Applying the Continuous Mapping Theorem (Theorem A.6) with the map $f : \bar{\mathbb{R}} \times \bar{\mathbb{R}} \to \mathbb{R} \times \mathbb{R}$ defined as $f(x, y) = (x - y, y)^T$ and set to an arbitrary value when $y = -\infty$,

$$\begin{bmatrix} \log \frac{p_{n,h'+h}}{p_{n,h'}} \\ \log \frac{p_{n,h'}}{p_{n,0}} \end{bmatrix} \quad \overset{P_{n,0}}{\rightsquigarrow} \quad \begin{bmatrix} \Delta_{h'+h} - \Delta_{h'} - \frac{1}{2} ||h||^2 - \langle h,h' \rangle \\ \Delta_{h'} - \frac{1}{2} ||h'||^2 \end{bmatrix} \equiv \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix},$$

where

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N \left(\begin{bmatrix} -\frac{1}{2} ||h||^2 - \langle h, h' \rangle \\ -\frac{1}{2} ||h'||^2 \end{bmatrix}, \begin{bmatrix} ||h||^2 & \langle h, h' \rangle \\ \langle h, h' \rangle & ||h'||^2 \end{bmatrix} \right).$$

Then, by Le Cam's Third Lemma (Theorem A.3),

$$\log \frac{p_{n,h'+h}}{p_{n,h'}} \quad \overset{P_{n,h'}}{\rightsquigarrow} \quad N\left(-\frac{1}{2}\left|\left|h\right|\right|^2, \left|\left|h\right|\right|^2\right),$$

which is the distribution of $\Delta_h - \frac{1}{2}\left|\left|h\right|\right|^2$.

The second claim follows directly from the above and the facts that $\left|\left|h(a)\right|\right|^2 = a^T\Sigma a$ and $\mathbb{E}((\Delta_{h(a)} - a^T\Delta)^2) = 0$ so that $\Delta_{h(a)} = a^T\Delta$ almost surely and therefore in law.

■

Lemma A.4 (Auxiliary Representation Lemma)**.** *Let X, Y, Z denote random elements defined on a probability space $(\Omega, \mathcal{F}, P)$, such that (X, Y) is independent of Z. Let g be a measurable real-valued function such that $g(X) \geqslant 0$ P-almost surely, and $\mathbb{E}(g(X)) = 1$. Then, there exist $\tilde{X}$ and $\tilde{Y}$ such that for any measurable function f and set C,*

$$\mathbb{E}(\mathbf{1}(f(X, Y, Z) \in C)g(X)) = \mathbb{E}(\mathbf{1}(f(\tilde{X}, \tilde{Y}, Z) \in C))$$

Proof. Define $\tilde{X}, \tilde{Y}$ to be be random variables with a probability distribution

$$L_{\tilde{X},\tilde{Y}}(A) = \mathbb{E}(\mathbf{1}_A(\tilde{X}, \tilde{Y})) = \mathbb{E}(\mathbf{1}_A(X, Y)g(X)).$$

The assumed properties of g ensure that $L_{\tilde{X},\tilde{Y}}$ is indeed a probability distribution, and since (X, Y) is independent of Z, $(\tilde{X}, \tilde{Y})$ is independent of Z as well. Next, define the function $f_z(x, y) = f(x, y, z)$ for every fixed z. Then,

$$\begin{aligned}
\mathbb{E}(\mathbf{1}(f(X, Y, Z) \in C)g(X)) &\overset{(a)}{=} \int \mathbb{E}(\mathbf{1}(f(X, Y, z) \in C)g(X))dP_Z(z) \\
&= \int \mathbb{E}(\mathbf{1}((X, Y) \in f_z^{-1}(C))g(X))dP_Z(z) \\
&\overset{(b)}{=} \int \mathbb{E}(\mathbf{1}((\tilde{X}, \tilde{Y}) \in f_z^{-1}(C)))dP_Z(z) \\
&= \int \mathbb{E}(\mathbf{1}(f(\tilde{X}, \tilde{Y}, z) \in C))dP_Z(z) \\
&\overset{(c)}{=} \mathbb{E}(\mathbf{1}(f(\tilde{X}, \tilde{Y}, Z) \in C))
\end{aligned}$$

where (a) follows from the independence of (X, Y) and Z, (b) holds by definition of

$(\tilde{X}, \tilde{Y})$, and (c) follows from the independence of $(\tilde{X}, \tilde{Y})$ and Z. ■

Lemma A.5 (Asymptotic Representation)**.** *Let $\phi_0' : \mathbb{B} \to \mathbb{D}$ denote an arbitrary map, $r, b_1, \ldots, b_m \in \mathbb{B}$ be arbitrary elements, Σ be a symmetric $m \times m$ matrix of full rank, $a \in \mathbb{R}^m$ be a vector and (V, Δ) be a tight random element in $\mathbb{D} \times \mathbb{R}^m$ with $\Delta \sim N(0, \Sigma)$ marginally Gaussian. Consider a measure on $\mathbb{D}$ given by*

$$L_a(C) = \mathbb{E}\left(\mathbf{1}_C\left(V - \phi_0'\left(r + \sum_{j=1}^{m} a_j b_j\right) + \phi_0'(r)\right) \exp\left\{a^T\Delta - \frac{1}{2}a^T\Sigma a\right\}\right)$$

Suppose that all of the elements introduced above are such that L_a defines a probability measure on $\mathbb{D}$ for all $a \in \mathbb{R}^m$. Define $S_\lambda = \lambda^{-1} I_m$ and $\Sigma_\lambda = (\Sigma + S_\lambda^{-1})^{-1}$ for any $\lambda > 0$. Then,

$$\int L_a dN(\mu, S_\lambda)(a) = \mathcal{L}\left(V_{\mu,\lambda,m} - \phi_0'(Z_{\lambda,m} + W_{\mu,\lambda,m} + r) + \phi_0'(r) \right)$$

as laws in $\mathbb{D}$, where $Z_{\lambda,m} = \sum_{j=1}^{m} p_j b_j$ for $p \sim N(0, \Sigma_\lambda)$ is a Gaussian random element in $\mathbb{B}$ independent from a tight random element $(V_{\mu,\lambda,m}, W_{\mu,\lambda,m})$ in $\mathbb{D} \times \mathbb{B}$.

Proof. Integrate both sides of the first display in the statement of the lemma over $a \sim N(\mu, S_\lambda)$. By Fubini's Theorem, the integrand of the right-hand side equals:

$$\mathbb{E}\Bigg(\mathbf{1}_C\left(V - \phi_0'\left(r + \textstyle\sum_{j=1}^{m} a_j b_j\right) + \phi_0'(r)\right) \frac{\det(S_\lambda)^{-1/2}}{(2\pi)^{m/2}} \\ \times \exp\left\{a^T\Delta - \frac{1}{2}a^T\Sigma a - \frac{1}{2}(a-\mu)^T S_\lambda^{-1}(a-\mu)\right\}\Bigg)$$

Observe that for $\Sigma_\lambda = (\Sigma + S_\lambda^{-1})^{-1}$ and $t = a - \mu - \Sigma_\lambda(\Delta - \Sigma\mu)$:

$$\exp\left\{a^T\Delta - \frac{1}{2}a^T\Sigma a - \frac{1}{2}(a-\mu)^T S_\lambda^{-1}(a-\mu)\right\} \cdot \frac{\det(S_\lambda)^{-1/2}}{(2\pi)^{m/2}} \\ = \frac{\det(\Sigma_\lambda)^{-1/2}}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}t^T\Sigma_\lambda^{-1}t\right\} \cdot c_{\mu,\lambda,m}(\Delta)$$

where

$$c_{\mu,\lambda,m}(\Delta) = \left(\frac{\det(\Sigma_\lambda)}{\det(S_\lambda)}\right)^{1/2} \cdot \exp\left\{\Delta^T\mu - \frac{1}{2}\mu^T\Sigma\mu + \frac{1}{2}(\Delta - \Sigma\mu)^T\Sigma_\lambda(\Delta - \Sigma\mu)\right\}$$

Denote $\omega = \mu + \Sigma_\lambda(\Delta - \Sigma\mu)$. Performing the change of variables described above, one obtains that $\int L_a(C)dN(\mu, S_\lambda)(a)$ equals

$$\int \mathbb{E}\left(\mathbf{1}_C\left(V - \phi_0'\left(r + \sum_{j=1}^m t_j b_j + \sum_{j=1}^m \omega_j b_j\right) + \phi_0'(r)\right) \cdot c_{\mu,\lambda,m}(\Delta)\right) dN(0, \Sigma_\lambda)(t). \tag{A.2}$$

Denote $Z_{\lambda,m} = \sum_{j=1}^m t_j b_j$. The law of $Z_{\lambda,m}$ in $\mathbb{B}$ is

$$L_{Z_{\lambda,m}}(B) = \int \mathbf{1}_B\left(\sum_{j=1}^m t_j b_j\right) dN(0, \Sigma_\lambda)(t).$$

Note that $c_{\mu,\lambda,m}(\Delta) \geqslant 0$ almost surely and $\mathbb{E}(c_{\mu,\lambda,m}(\Delta)) = 1$ by construction. Let $(V_{\mu,\lambda,m}, W_{\mu,\lambda,m})$ be a tight random element in $\mathbb{D} \times \mathbb{B}$, with the law

$$L_{(V_{\mu,\lambda,m}, W_{\mu,\lambda,m})}(A) = \mathbb{E}\left(\mathbf{1}_A(V_{\mu,\lambda,m}, W_{\mu,\lambda,m})\right) \equiv \mathbb{E}\left(\mathbf{1}_A\left\{\left(V, \sum_{j=1}^m \omega_j b_j\right)\right\} \cdot c_{\mu,\lambda,m}(\Delta)\right).$$

Apply Lemma A.4 with $X = (X_1, X_2) = (\sum_{j=1}^m \omega_j b_j, \Delta)$, $Y = V$, $\tilde{X} = W_{\mu,\lambda,m}$, $\tilde{Y} = V_{\mu,\lambda,m}$, and $Z = Z_{\lambda,m}$, and the maps

$$\begin{aligned} f(x,y,z) &= y - \phi_0'(r + z + x_1) + \phi_0'(r) \\ g(x) &= c_{\mu,\lambda,m}(x_2) \end{aligned}$$

It follows that Equation (A.2) represents the law of

$$Q_{\mu,\lambda,m} \equiv V_{\mu,\lambda,m} - \phi_0'(Z_{\lambda,m} + W_{\mu,\lambda,m} + r) + \phi_0'(r)$$

and the proof is complete.

■

Lemma A.6 (Asymptotic Tightness and Asymptotic Measurability of a Shifted Sequence)**.** *Let $X_n : \Omega_n \to \mathbb{D}$ be a sequence of maps into a Banach space $(\mathbb{D}, ||\cdot||_{\mathbb{D}})$. Assume that X_n is asymptotically tight and asymptotically measurable. Let $(c_n)_{n\geq 1} \in \mathbb{D}$ denote a sequence of constants such that $||c_n - c||_{\mathbb{D}} \to 0$ for some $c \in \mathbb{D}$. Then $X_n + c_n$ is asymptotically tight and asymptotically measurable.*

Proof. Since X_n is asymptotically tight, for every $\varepsilon > 0$ there is a compact set K_ε such that $\liminf_{n\to\infty} P_*(X_n \in K_\varepsilon^\delta) \geqslant 1-\varepsilon$ for every $\delta > 0$, where K_ε^δ is the set of points within distance δ from K_ε. Clearly, $X_n + c$ is asymptotically tight since it satisfies the inequality above with $\tilde{K}_\varepsilon = K_\varepsilon + c$ in place of K_ε. Next, fix $\varepsilon > 0$ and $\delta > 0$. There is n_0 such that for all $n \geqslant n_0$, $d(c_n, c) < \delta/2$. Then, for any compact K, $X_n + c \in K^{\delta/2}$ implies that $X_n + c_n \in K^\delta$ for all $n \geqslant n_0$. Therefore:

$$\liminf_{n\to\infty} P_*(X_n + c_n \in \tilde{K}_\varepsilon^\delta) \geqslant \liminf_{n\to\infty} P_*(X_n + c \in \tilde{K}_\varepsilon^{\delta/2}) \geqslant 1-\varepsilon$$

so that $X_n + c_n$ is tight.

Next, since X_n and c_n are marginally asymptotically measurable and asymptotically tight, they are jointly asymptotically measurable (Lemma 1.4.4. in van der Vaart and Wellner, 1996). Therefore, for any $f \in C_b(\mathbb{D} \times \mathbb{D})$ it holds that $\mathbb{E}^*(f(X_n, c_n)) - \mathbb{E}_*(f(X_n, c_n)) \to 0$ as $n \to \infty$. In particular, this holds for $f(X_n, c_n) = g(X_n + c_n)$ for any $g \in C_b(\mathbb{D})$, which implies the asymptotic measurability of $X_n + c_n$. ■

Lemma A.7 (Lipchitzness of the Asymptotic Risk)**.** *Let l_M be a loss function satisfying Remark 1, and ϕ be a directionally differentiable function satisfying Assumption 2.1. Let $(\mathbb{D}, ||\cdot||_\mathbb{D})$ and $(\mathbb{B}, ||\cdot||_\mathbb{B})$ denote Banach spaces and Z denote a tight random element in $\mathbb{B}$. Then a function $f : \mathbb{D} \times \mathbb{B} \times \mathbb{B} \to \mathbb{R}$ defined as*

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

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

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

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

■

Lemma A.8 (Approximating Sub-Convex Loss Functions)**.** *Any subconvex loss function l (see Assumption 3.3) can be approximated by a sequence of bounded Lipschitz functions l_M pointwise monotonically from below.*

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

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

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

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

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

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

First, let $i = j$. Then

$$\begin{aligned}
|l_{M,r}(y) - l_{M,r}(x)| &= \frac{1}{2^r}\left|\frac{d(y,B_i)d(x,F_{M,i}) - d(x,B_i)d(y,F_{M,i})}{(d(y,B_i)+d(y,F_{M,i}))(d(x,B_i)+d(x,F_{M,i}))}\right| \\
&= \frac{1}{2^r}\left|\frac{d(y,B_i)(d(x,F_{M,i}) - d(y,F_{M,i})) + d(y,F_{M,i})(d(y,B_i) - d(x,B_i))}{(d(y,B_i)+d(y,F_{M,i}))(d(x,B_i)+d(x,F_{M,i}))}\right| \\
&\overset{\text{(a)}}{\leqslant} \frac{1}{2^r}\cdot\frac{(d(y,B_i)+d(y,F_{M,i}))\cdot d(x,y)}{(d(y,B_i)+d(y,F_{M,i}))(d(x,B_i)+d(x,F_{M,i}))} \\
&\overset{\text{(b)}}{\leqslant} \frac{M}{2^r}\cdot d(x,y) \qquad \text{(A.3)}
\end{aligned}$$

Where (a) follows from the reverse triangle inequality, i.e. $|d(y, B_i) - d(x, B_i)| \leqslant$

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

Proof of Theorem 1. While the theorem is stated assuming that the tangent set $T(P)$ is a linear space, the proof below covers a more general case, allowing $T(P)$ to be a convex cone. To accommodate convex cones, $T(P)$ is replaced with $\overline{\text{lin}}\, T(P)$ in the Definition 3.3, so that the path-wise derivative $\theta_0'(h)$ is defined on $\overline{\text{lin}}\, T(P)$. The lower bound holds as stated in the Theorem, with $S(Z)$ replaced by $\theta_0'(T(P))$.

The proof consists of three main steps. The first step is to establish the weak limits (along subsequences) of an estimator $\sqrt{n}(\hat{\phi}_n - \phi(\theta_0)) \rightsquigarrow L_{a,h}$ along $P_{n,h+h(a)}$ where $h \in T(P)$ and $h(a) = \sum_{j=1}^m a_j h_j$ with $a_1,\dots,a_m \in \mathbb{R}$ and linearly independent $h_1,\dots,h_m \in T(P)$. This is made possible by the asymptotic normality of the log-likelihood ratios (Lemma A.3), the assumed differentiability of $\theta(P)$, and Le Cam's Third Lemma.

The second step is to show that a suitable average of the limiting laws $L_{a,h}$, over a, for a fixed h, can be represented as the law of $V - \phi_0'(Z_m + W + \theta_0'(h)) + \phi_0'(\theta_0'(h))$, where Z_m is a sequence of Gaussian random elements, which does not depend on h, and (V,W), independent of Z_m, represent some "noise terms" with unknown distributions, which may depend on h. This representation allows to bound the LAM risk of a particular sequence $\hat{\phi}_n$ from below by $\sup_{h \in T(P)} \mathbb{E}(l(V - \phi_0'(Z_m + W + \theta_0'(h)) + \phi_0'(\theta_0'(h))))$. Different estimators $\hat{\phi}_n$ correspond to different distributions of (V,W), so, to obtain a lower bound that would hold for all estimators, I take the infimum over all possible distributions of (V,W).

The final step is to show that the infimum is attained by constants (i.e., degenerate distributions). This is made possible by the general purification result of Feinberg and Piunovskiy (2006) on matching randomized decision rules with the non-randomized ones. Combining the result with the assumed symmetry of the loss function, letting m approach infinity, and taking care of technical details yields the final form of the lower bound.

I start with some preliminary notation. Let $\bar{h} = (h_1,\dots,h_m)$ denote the first m elements of a linearly independent sequence in $T(P)$, and h denote an arbitrary element of $T(P)$. Let $\Sigma = \mathbb{E}(\bar{h}\bar{h}^T)$ be a $m \times m$ scalar-product matrix. Denote

$h(a) = \sum_{j=1}^{m} a_j h_j$ for an arbitrary $a = (a_1, \ldots, a_m) \in \mathbb{R}^m$. Since $T(P)$ is a convex cone, for any $a \in \mathbb{R}^m_+$, it holds that $h + h(a) \in T(P)$. On the other hand, for $a \notin \mathbb{R}^m_+$ it could be that $h + h(a) \notin T(P)$.

Consider a sequence of random elements $\{Z_m\} \in \mathbb{B}$ where $Z_m = \sum_{j=1}^{m} a_j \theta_0'(h_j)$, where $a \sim N(0, \Sigma^{-1})$. It follows from the proof of Theorem 3.11.2 in van der Vaart and Wellner (1996) that this sequence is uniformly tight. Next, let $b_1^*, \ldots, b_K^* \in \mathbb{B}^*$ denote arbitrary elements of the continuous dual of $\mathbb{B}$. Each $b_k^* \circ \theta_0'$ is a bounded linear map from $\overline{\text{lin}}\ T(P)$ into $\mathbb{R}$. By the Riesz Representation Theorem, there is an element $\tilde{\theta}_k \in \overline{\text{lin}}\ T(P)$ such that: $b_k^* \circ \theta_0'(h) = \mathbb{E}_P(\tilde{\theta}_k h)$ for any $h \in \overline{\text{lin}}\ T(P)$. Let $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_K)^T$ and B denote a $K \times m$ matrix with $B_{i,j} = \mathbb{E}_P(\tilde{\theta}_i h_j)$ for $i = 1, \ldots, K$ and $j = 1, \ldots, m$. Then $\tilde{\theta}^m = B\Sigma^{-1}\bar{h}$ is the orthogonal projection of $\tilde{\theta}$ onto the closed linear span of $(h_1, \ldots, h_m)$. By construction, $\tilde{\theta}^m \to \tilde{\theta}$ in $L_2(P)$ as $m \to \infty$, implying that $\mathbb{E}(\tilde{\theta}^m(\tilde{\theta}^m)^T) = B\Sigma B^T \to \mathbb{E}_P(\tilde{\theta}\tilde{\theta}^T)$ in Frobenius norm. It follows that $(b_1^*, \ldots, b_K^*) \circ Z_m \rightsquigarrow N(0, \mathbb{E}(\tilde{\theta}\tilde{\theta}^T))$ as $m \to \infty$. Since $\{Z_m\}$ is asymptotically tight, and the weak limits along every its subsequence are the same, it follows that $Z_m \rightsquigarrow Z$ in $\mathbb{B}$ where Z is a tight Gaussian process such that $(b_1^*, \ldots, b_K^*) \circ Z$ is a centered Normal random vector with covariance $\mathbb{E}_P(\tilde{\theta}\tilde{\theta}^T)$.

Step 1: Subsequence Argument

Define $\Delta_{n,h}$ and Δ as in Lemma A.3. Pick an arbitrary $h \in T(P)$, recall that Definition 3.1 implies that, for $P_{n,0} = \prod_{i=1}^{n} P$ and $P_{n,h} = \prod_{i=1}^{n} P_{1/\sqrt{n},h}$,

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

where $\Delta_{n,h} \rightsquigarrow_{P_{n,0}} \Delta_h$ with $\Delta_h \sim N(0, ||h||^2)$.

Let $\hat{\phi}_n$ denote an arbitrary estimator sequence such that $\sqrt{n}(\hat{\phi}_n - \phi(\theta_0))$ is asymptotically tight and asymptotically measurable under $P_{n,0}$. By the Prohorov's Theorem (Theorem A.7) and the General Le Cam's Third Lemma (Theorem A.2), for any subsequence, there is a further subsequence, still denoted by $\{n\}$ for simplicity, such that $\sqrt{n}(\hat{\phi}_n - \phi(\theta_0))$ converges weakly to a tight limit under $P_{n,h}$. Lemma 1.3.8. in van der Vaart and Wellner (1996) then implies that the subsequence $\sqrt{n}(\hat{\phi}_n - \phi(\theta_0))$ is asymptotically tight and asymptotically measurable under $P_{n,h}$.

Denoting $\theta_n(h) = \theta(P_{n,h})$, by the differentiability of $\theta(P)$, one may write $\theta_n(h) =$

$\theta_0+\theta_0'(h)/\sqrt{n}+o(1/\sqrt{n})$. Then, by the directional differentiability of ϕ, $\sqrt{n}(\phi(\theta_n(h))-\phi(\theta_0)) \to \phi_0'(\theta_0'(h))$ as $n \to \infty$. Therefore, the seqence

$$\sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h)) = \sqrt{n}(\hat{\phi}_n - \phi(\theta_0)) - \sqrt{n}(\phi(\theta_n(h)) - \phi(\theta_0))$$

is asymptotically tight and asymptotically measurable under $P_{n,h}$ by Lemma A.6. By Lemma A.3 and Prohorov's theorem (Theorem A.7), there is a further subsequence such that, for a random element $V \in \mathbb{D}$ with a tight Borel law,

$$\begin{bmatrix} \sqrt{n}\left(\hat{\phi}_n - \phi(\theta_n(h))\right) \\ \log \frac{dP_{n,h+h(a)}}{dP_{n,h}} \end{bmatrix} \overset{P_{n,h}}{\rightsquigarrow} \begin{bmatrix} V \\ a'\Delta - \frac{1}{2}a^T\Sigma a \end{bmatrix}$$

in $\mathbb{D} \times \mathbb{R}$, for any $h \in T(P)$ and any $a \in \mathbb{R}^m$ such that $h + h(a) \in T(P)$. The law of V depends on h, although obviated from the notation for now. Note that:

$$\begin{aligned} \sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h + h(a)))) \;&= \sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h))) \\ &- \sqrt{n}(\phi(\theta_n(h + h(a))) - \phi(\theta_0)) \\ &+ \sqrt{n}(\phi(\theta_n(h)) - \phi(\theta_0)). \end{aligned}$$

By the assumed differentiability of $\theta(P)$ and directional differentiability of ϕ,

$$\begin{aligned} \sqrt{n}(\phi(\theta_n(h + h(a))) - \phi(\theta_0)) \;&\to\; \phi_0'\left(\theta_0'(h) + \sum_{j=1}^{m} a_j\theta_0'(h_j)\right) \\ \sqrt{n}(\phi(\theta_n(h)) - \phi(\theta_0)) \;&\to\; \phi_0'\left(\,\theta_0'(h)\,\right) \end{aligned}$$

in $\mathbb{D}$, as $n \to \infty$. By Slutsky's Lemma, for $a \in \mathbb{R}^m$ such that $h + h(a) \in T(P)$,

$$\begin{bmatrix} \sqrt{n}\left(\hat{\phi}_n - \phi(\theta_n(h + h(a)))\right) \\ \log \frac{dP_{n,h+h(a)}}{dP_{n,h}} \end{bmatrix} \overset{P_{n,h}}{\rightsquigarrow} \begin{bmatrix} V - \phi_0'\left(\theta_0'(h) + \sum_{j=1}^{m} a_j\theta_0'(h_j)\right) + \phi_0'\left(\,\theta_0'(h)\,\right) \\ a'\Delta - \frac{1}{2}a^T\Sigma a \end{bmatrix}$$

in $\mathbb{D} \times \mathbb{R}$. The General Le Cam's Third Lemma (Theorem A.3) implies that, for

$h + h(a) \in T(P)$,

$$\sqrt{n}\left(\hat{\phi}_n - \phi(\theta_n(h + h(a))\right) \quad \overset{P_{n,h+h(a)}}{\rightsquigarrow} \quad L_{a,h}$$

in $\mathbb{D}$, where for any Borel $C \in \mathbb{D}$, $L_{a,h}(C)$ is given by

$$\mathbb{E}\left(\mathbf{1}_C\left(V - \phi_0'\left(\theta_0'(h) + \sum_{j=1}^m a_j\theta_0'(h_j)\right) + \phi_0'\left(\theta_0'(h)\right)\right)\exp\left\{a'\Delta - \frac{1}{2}a^T\Sigma a\right\}\right),$$

which defines a tight probability measure on $\mathbb{D}$. Note, however, that $L_{a,h}$ defines a tight probability measure *for any* $h \in T(P)$ *and any* $a \in \mathbb{R}^m$. Indeed, by direct computation, $L_{a,h}(\mathbb{D}) = \mathbb{E}\left(\exp\{a'\Delta - \frac{1}{2}a^T\Sigma a\}\right) = 1$ and the tightness of $L_{a,h}$ follows immediately from the tightness of $V - \phi_0'(\theta_0'(h) + \sum_{j=1}^m a_j\theta_0'(h_j)) + \phi_0'(\theta_0'(h))$.

Step 2: Main Argument

Let R denote the local asymptotic maximum risk (i.e. the left-hand side) in the statement of the theorem. Direct the finite subsets of the tangent set by inclusion. There exists a subnet $\{n_I : I \subset T(P), I \text{ is finite}\}$ such that

$$R = \lim_I \sup_{h\in I} \mathbb{E}_{*P_{n_I,h}}\left\{l\left(\sqrt{n}(\hat{\phi}_{n_I} - \phi(\theta_{n_I}(h)))\right)\right\}. \tag{A.4}$$

By the preceding argument,[27] for any $h \in T(P)$ and $a \in \mathbb{R}^m$ such that $h+h(a) \in T(P)$ there is a further subsequence, still denoted by $\{n\}$ for simplicity, along which

$$\sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h + h(a)))) \quad \overset{P_{n,h+h(a)}}{\rightsquigarrow} \quad L_{a,h}$$

in $\mathbb{D}$. Let l_M denote a loss function satisfying Remark 1. Then, for any $h \in T(P)$ and any $a \in \mathbb{R}^m$ such that $h + h(a) \in T(P)$,

$$R \geqslant \liminf_{n\to\infty} \mathbb{E}_{*P_{n,h+h(a)}}\left\{l_M\left(\sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h + h(a))))\right)\right\} \geqslant \int l_M dL_{a,h},$$

where the first inequality is due to fixing a single $h + h(a) \in T(P)$, and switching to $l_M \leqslant l$, and the second inequality follows from the Portmanteau Theorem (Theorem

[27] The argument was presented for subsequences, but it holds without modifications for subnets as well. I will use the term subsequence in the sequel.

1.3.4 in van der Vaart and Wellner, 1996).

Denote $S_\lambda = \lambda^{-1} I_m$ and $\Sigma_\lambda = (\Sigma + S_\lambda^{-1})^{-1}$ for some $\lambda > 0$. Lemma A.5 shows that:

$$\int L_{a,h} dN(\mu, S_\lambda)(a) = \mathcal{L}\left(V_{\mu,\lambda,m} - \phi_0'(Z_{\lambda,m} + W_{\mu,\lambda,m} + \theta_0'(h)) + \phi_0'(\theta_0'(h)) \right)$$

as laws in $\mathbb{D}$, where $Z_{\lambda,m} = \sum_{j=1}^m p_j \theta_0'(h_j)$, with $p \sim N(0, \Sigma_\lambda)$, is a Gaussian element in $\mathbb{B}$, and $(V_{\mu,\lambda,m}, W_{\mu,\lambda,m})$ a tight random element in $\mathbb{D} \times \mathbb{B}$, independent of $Z_{\lambda,m}$. In what follows the subscripts will be obviated to simplify the notation.

Integrating both sides of the inequality $R \geqslant \int l_M dL_{a,h}$ with respect to $dN(\mu, S_\lambda)(a)$ over $a \in \mathbb{R}^m$ such that $h + h(a) \in T(P)$ yields

$$\begin{aligned} R \geqslant \int \left(\int l_M dL_{a,h} \right) \mathbf{1}\{h + h(a) \in T(P)\} dN(\mu, S_\lambda)(a) \\ \geqslant \mathbb{E}\left\{ l_M\left(V - \phi_0'\left(Z + W + \theta_0'(h)\right) + \phi_0'\left(\theta_0'(h)\right)\right)\right\} \\ - B_M \int \mathbf{1}\{h + h(a) \notin T(P)\} dN(\mu, S_\lambda)(a). \quad \text{(A.5)} \end{aligned}$$

Since $L_2(P)$ is separable, there is a countable dense subset of $T(P)$, denoted T_0. Since the inequalities in the above display hold for any $h \in T(P)$, the LAM risk R is bounded from below by

$$\begin{aligned} \sup_{h \in T_0} \int \mathbb{E}_Z \left\{ l_M\left(v - \phi_0'\left(Z + w + \theta_0'(h)\right) + \phi_0'\left(\theta_0'(h)\right)\right)\right\} dF(v, w; h) \\ - B_M \sup_{h \in T_0} \int \mathbf{1}\{h + h(a) \notin T(P)\} dN(\mu, S_\lambda)(a). \quad \text{(A.6)} \end{aligned}$$

Consider the second summand. Note that, for any $h \in T(P)$, $\{h + h(a) \notin T(P)\}$ implies $\{h(a) \notin T(P)\}$ which further implies $\{a \notin \mathbb{R}_+^m\}$ because $T(P)$ is a convex cone. Therefore:

$$\begin{aligned} \sup_{h \in T_0} \int \mathbf{1}\{h + h(a) \notin T(P)\} dN(\mu, S_\lambda)(a) &\leqslant \int \mathbf{1}\{h(a) \notin T(P)\} dN(\mu, S_\lambda)(a) \\ &\leqslant \int \mathbf{1}\{a \notin \mathbb{R}_+^m\} dN(\mu, S_\lambda)(a). \end{aligned}$$

Let $\mu = \delta \cdot (1, \ldots, 1)^T \in \mathbb{R}^m_+$, apply the above inequality and pass to the limit in A.6 as $\delta \uparrow \infty$. Then,

$$\begin{aligned} R &\geqslant \sup_{h \in T_0} \mathbb{E}\left\{ l_M \left(V - \phi_0' \left(Z + W + \theta_0'(h) \right) + \phi_0'(\theta_0'(h)) \right) \right\} \\ &= \sup_{h \in T_0} \int \mathbb{E}_Z \left\{ l_M \left(v - \phi_0' \left(Z + w + \theta_0'(h) \right) + \phi_0' \left(\theta_0'(h) \right) \right) \right\} dQ_h(v, w), \end{aligned} \tag{A.7}$$

where the equality follows from the independence of Z and (V, W) and $Q_h(v, w)$ denotes the joint distribution of the latter, explicitly indexed by h.

Step 3: Purification Argument

The lower bound in Equation (A.7) can be tightened by taking an infimum over all probability measures $Q_h(v, w)$, on $\mathbb{D} \times \mathbb{B}$, but the result would not be practically useful. To this end, following the idea from Song (2014) and Fang (2018), I employ a purification technique, in this case by Feinberg and Piunovskiy (2006), as described in Theorem A.8.

Since the space $\mathbb{D} \times \mathbb{B}$ may not be separable, to apply the aforementioned result, I use a compactification step. Enumerate the elements of $T_0 = \{h_1, h_2, \ldots\}$, and denote $r_j = \theta_0'(h_j)$ for $j \geqslant 1$. Recall that, for each $h_j \in T(P)$, the distribution $Q_{h_j}(v, w)$ is tight, that is, for each $\varepsilon > 0$ there is a compact set $A_j \subset \mathbb{D} \times \mathbb{B}$ such that $Q_{h_j}(A_j) \geqslant 1 - \varepsilon$. Fix some $\varepsilon > 0$, and $J \in \mathbb{N}$, and let $A = \cup_{j=1}^J A_j \subset \mathbb{D} \times \mathbb{B}$ be a compact set such that $Q_{h_j}(A) \geqslant 1 - \varepsilon$ for all $j = 1, \ldots, J$. Define the distributions $\tilde{Q}_j$ supported on A as

$$\tilde{Q}_j(B) \equiv \frac{Q_{h_j}(B \cap A)}{Q_{h_j}(A)}$$

for $j = 1, \ldots, J$, for any Borel B. Then, for any non-negative measurable function $f : \mathbb{D} \times \mathbb{B} \to \mathbb{R}$,

$$\int f(v, w) dQ_{h_j}(v, w) \geqslant (1 - \varepsilon) \int_A f(v, w) d\tilde{Q}_j(v, w). \tag{A.8}$$

In the notation of Theorem A.8, let $X = [0, J]$ denote the state space, and the set A, constructed above, denote the action space, both endowed with the corresponding Borel sigma-fields. Consider the uniform distributions $\mu_k = U[k - 1, k]$ for $k =$

$1, \dots, J$ on X, and a randomized decision rule $\pi : X \to \mathcal{P}(A)$ defined as

$$\pi(v, w; x) = \tilde{Q}_j(v, w) \quad \text{if } x \in [j-1, j]\,, \quad \text{for } j = 1, \dots, J$$

for $(v, w) \in \mathbb{D} \times \mathbb{B}$. Finally, define the loss functions, for $j = 1, \dots, J$,

$$\rho_j(v, w, x) = \mathbb{E}_Z\left(l_M\left(v - \phi_0'\left(Z + w + r_j\right) + \phi_0'(r_j)\right)\right).$$

With this notation,

$$\begin{aligned}\int_A \mathbb{E}_Z\left(l_M\left(v - \phi_0'(Z + w + r_j) + \phi_0'(r_j)\right)\right) d\tilde{Q}_j(v, w)& \\ = \int\limits_X \int\limits_A \rho_j((v, w), x) d\pi((v, w); x) d\mu_j(x)& \qquad \text{(A.9)}\end{aligned}$$

and, by Theorem A.8, there exists a measurable map $(v, w) : X \to A$ such that

$$\int\limits_X \int\limits_A \rho_j((v, w), x) d\pi((v, w); x) d\mu_k(x) = \int\limits_X \rho_j((v(x), w(x)), x) d\mu_k(x) \qquad \text{(A.10)}$$

for all $j = 1, \dots, J$ and $k = 1, \dots, J$. In particular, the equality holds for all $j = k = 1, \dots, J$. Therefore:

$$\begin{aligned}&\max_{j \leqslant J} \int \mathbb{E}_Z\left(l_M\left(v - \phi_0'(Z + w + r_j) + \phi_0'(r_j)\right)\right) dQ_{h_j}(v, w) \\ &\quad\overset{\text{(a)}}{\geqslant} (1 - \varepsilon) \max_{j \leqslant J} \int_A \mathbb{E}_Z\left(l_M\left(v - \phi_0'(Z + w + r_j) + \phi_0'(r_j)\right)\right) d\tilde{Q}_j(v, w) \\ &\quad\overset{\text{(b)}}{=} (1 - \varepsilon) \max_{j \leqslant J} \int\limits_{j-1}^{j} \mathbb{E}_Z\left(l_M\left(v(x) - \phi_0'(Z + w(x) + r_j) + \phi_0'(r_j)\right)\right) dx \\ &\quad\overset{\text{(c)}}{\geqslant} (1 - \varepsilon) \inf_{(v,w) \in \mathbb{D} \times \mathbb{B}} \max_{j \leqslant J} \int\limits_{j-1}^{j} \mathbb{E}_Z\left(l_M\left(v - \phi_0'(Z + w + r_j) + \phi_0'(r_j)\right)\right) dx \\ &\quad\overset{\text{(d)}}{=} (1 - \varepsilon) \inf_{(v,w) \in \mathbb{D} \times \mathbb{B}} \max_{j \leqslant J} \mathbb{E}_Z\Big(l_M\left(v - \phi_0'(Z + w + r_j) + \phi_0'(r_j)\right)\Big) \qquad \text{(A.11)}\end{aligned}$$

where (a) follows from Equation (A.8), (b) from Equations (A.9)–(A.10) and the definitions of ρ_j and μ_j, (c) is due to the fact that $(v(x), w(x))$ belongs to the range of (v,w) as x varies, and the range is included in $\mathbb{D}\times\mathbb{B}$, and (d) is due to the fact that the integrand does not depend on x. Recall that $Z = Z_{\lambda,m}$ and assume that the infimum in the display below can be attained, so that it can be equivalently taken over sufficiently large compact set. Letting $\varepsilon \to 0$ and $J \to \infty$ in (A.11) (with the help of Dini's Theorem) and recalling (A.7) yields, the lower bound:

$$R \ \geqslant \inf_{(v,w)\in\mathbb{D}\times\mathbb{B}} \sup_{h\in T_0} \mathbb{E}_Z \left(l_M(v - \phi_0'(Z_{\lambda,m} + w + \theta_0'(h)) + \phi_0'(\theta_0'(h)) \right).$$

Since the expectation on the preceding display is continuous in h, the set T_0 can be replaced with $T(P)$, by Lemma A.1. Writing the supremum over $s \in \theta_0'(T(P))$,

$$R \geqslant \inf_{(v,w)\in\mathbb{D}\times\mathbb{B}} \sup_{s\in\theta_0'(T(P))} \mathbb{E}_Z \left(l_M \left(v - \phi_0'\left(Z_{\lambda,m} + w + s\right) + \phi_0'\left(s\right)\right) \right). \tag{A.12}$$

Back To The Main Argument

Recall that $Z_{\lambda,m} = \sum_{j=1}^{m} a_j \theta_0'(h_j)$ with $a \sim N(0, \Sigma_\lambda)$ where $\Sigma_\lambda = (\Sigma + \lambda I)^{-1}$. As $\lambda \to 0$ and then $m \to \infty$, $Z_{\lambda,m} \rightsquigarrow Z$ as random elements in $\mathbb{B}$, where Z is a tight centered Gaussian process for which $(b_1^*, \ldots, b_K^*) \circ Z$ is a centered Normal random vector with covariance $\mathbb{E}_P(\tilde{\theta}\tilde{\theta}^T)$.[28] Pick a subsequence, indexed by l, along which $Z_l \equiv Z_{\lambda(l),m(l)} \rightsquigarrow Z$.

Weak convergence of Z_l to Z can be characterized by the point-wise convergence of linear operators $L_l(g) \equiv \mathbb{E}(g(Z_l)) \to \mathbb{E}(g(Z)) \equiv L_l(g)$ for all non-negative bounded Lipchitz functions g (Theorem 1.3.4 in van der Vaart and Wellner, 1996). However, this point-wise convergence is automatically uniform over certain subsets of this class. In particular, let $\mathcal{G}$ denote the set of all non-negative functions bounded by C_1 with Lipschitz constant C_2, that is,

$$\mathcal{G}(C_1, C_2) = \Big\{ g : \mathbb{B} \to \mathbb{R}_+ : \sup_z |g(z)| \leqslant C_1, \\ |g(z_1) - g(z_2)| \leqslant C_2 \left|\left|z_1 - z_2\right|\right|_{\mathbb{B}} \text{ for all } z_1, z_2 \in \mathbb{B} \Big\}.$$

[28] Since the limit law is fully characterized by the marginals, the joint weak convergence of marginals and asymptotic tightness suffices to deduce weak convergence in $\mathbb{B}$. See the discussion preceding the Subsequence Argument.

Then, by Theorem 1.12.1 in van der Vaart and Wellner (1996),

$$\sup_{g \in \mathcal{G}(C_1, C_2)} |\mathbb{E}(g(Z_l)) - \mathbb{E}(g(Z))| \to 0 \text{ as } l \to \infty.$$

Let $s \in \theta_0'(T(P))$. Note that, by Assumptions 2.1 and 3.3, the functions $g(\cdot; v, w, s) : \mathbb{B} \to \mathbb{R}$, defined as

$$g(z; v, w, s) = l_M\left(v - \phi_0'(z + w + s) + \phi_0'(s)\right),$$

are uniformly, in (z, v, w, s), bounded by B_M and uniformly, in (v, w, s), Lipchitz continuous in z with Lipschitz constant $C_M \cdot C_{\phi'}$. Indeed, for an arbitrary tuple $(v, w, s) \in \mathbb{D} \times \mathbb{B} \times \mathbb{B}$ and arbitrary $z, z' \in \mathbb{B}$,

$$\begin{aligned} |g(z; v, w, s) - g(z'; v, w, s)| &\leqslant C_M \cdot ||\phi_0'(z + w + s) - \phi_0'(z' + w + s)||_{\mathbb{D}} \\ &\leqslant C_M \cdot C_{\phi'} \cdot ||z - z'||_{\mathbb{B}}. \end{aligned}$$

Therefore, the class of functions

$$\tilde{\mathcal{G}} = \{g(z; v, w, s) : (v, w, s) \in \mathbb{D} \times \mathbb{B} \times \mathbb{B}\}$$

is a subset of $\mathcal{G}(B_M, C_M \cdot C_{\phi'})$, so that

$$\sup_{(v,w,s) \in \mathbb{D} \times \mathbb{B}^2} \Big| \mathbb{E}(g(Z_l; v, w, s)) - \mathbb{E}(g(Z; v, w, s)) \Big| \to 0.$$

Letting l approach infinity in Equation A.12, so that $\lambda(l) \to 0$ and $m(l) \to \infty$, yields:

$$R \geqslant \inf_{(v,w) \in \mathbb{D} \times \mathbb{B}} \sup_{s \in \theta_0'(T(P))} \mathbb{E}_Z\left(l_M\left(v - \phi_0'(Z + w + s) + \phi_0'(s)\right)\right).$$

Note that, by construction, Z concentrates on $S(Z) = \overline{\text{lin}}\, \theta_0'(T(P))$. If $T(P)$ itself is a linear space, $S(Z) = \overline{\text{lin}}\, \theta_0'(T(P)) = \overline{\theta_0'(T(P))}$. In this case, since the expectation in the above display is continuous in s, the supremum can be equivalently taken over $S(Z)$ by Lemma A.1. Then, the proof can be continued with $S(Z)$ in place of $\theta_0'(T(P))$.

Define:

$$f(v,w) = \sup_{s\in\theta_0'(T(P))} \mathbb{E}_Z(l\left(v - \phi_0'(Z+w+s) + \phi_0'(s)\right))$$

and

$$f_M(v,w) = \sup_{s\in\theta_0'(T(P))} \mathbb{E}_Z(l_M\left(v - \phi_0'(Z+w+s) + \phi_0'(s)\right)),$$

and assume that the infimum of $f(v,w)$ can be attained. Let K denote a large enough compact set that contains at lease one minimizer of f. I will argue that $f_M(v,w)$ is a sequence of continuous functions converging point-wise monotonically to a continuous function $f(v,w)$ for all $(v,w) \in K$. First, by Assumption 3.3, for each M, $f_M(v,w) \leqslant f(v,w)$ so that $\limsup_{M\to\infty} f_M(v,w) \leqslant f(v,w)$. On the other hand,

$$\begin{aligned}
\liminf_{M\to\infty} f_M(v,w) &= \liminf_{M\to\infty} \sup_{s\in\theta_0'(T(P))} \mathbb{E}_Z\{l_M(\phi_0'(Z+w+s) - \phi_0'(s) + v)\} \\
&\geqslant \sup_{s\in\theta_0'(T(P))} \liminf_{M\to\infty} \mathbb{E}_Z\{l_M(\phi_0'(Z+w+s) - \phi_0'(s) + v)\} \\
&\overset{(a)}{\geqslant} \sup_{s\in\theta_0'(T(P))} \mathbb{E}_Z\{l(\phi_0'(Z+w+s) - \phi_0'(s) + v)\} \\
&= f(v,w),
\end{aligned}$$

where (a) follows from Fatou's Lemma and the assumed point-wise convergence of l_M. Therefore, $f_M(v,w)$ converges point-wise to $f(v,w)$. Then, by Dini's Theorem, this convergence is also uniform over K, which completes the proof of the Theorem.

Remark 1 is proven similarly, by defining R_M as R in equation (A.4) with l_M instead of l.

■

Proof of Theorem 2. Consider an estimator sequence of the form

$$\hat{\phi}_n = \phi\left(\hat{\theta}_n + \frac{\hat{v}_{1,n}}{\sqrt{n}}\right) + \frac{\hat{v}_{2,n}}{\sqrt{n}}, \tag{A.13}$$

where $\hat{\theta}_n$ is the best regular estimator for θ_0 in the sense of the Convolution Theorem (A.5), and $\hat{v}_{1,n}$, $\hat{v}_{2,n}$ are adjustment terms depending on the data. To calculate the

LAM risk of this estimator sequence, it is necessary to study its distributional limits under the "local perturbations" $P_{n,h}$ (see Definition 3.4).

Let $v_1 \in \mathbb{B}$ and $v_2 \in \mathbb{D}$ denote the probability limits of $\hat{v}_{1,n}$ and $\hat{v}_{2,n}$ under $P_{n,h}$ correspondingly.[29] Since $\hat{\theta}_n$ is best regular, $\sqrt{n}(\hat{\theta}_n - \theta(P_{n,h})) \rightsquigarrow_{P_{n,h}} \mathbb{G}_0$, where the limit distribution is the same for any $h \in T(P)$. Since $\theta(P)$ is differentiable, $\sqrt{n}(\theta(P_{n,h}) - \theta_0) = \theta_0'(h)$ for any $h \in T(P)$. Then, by the Slutsky's Theorem,

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

as random elements in $\mathbb{B}$. The assumed differentiability of $\theta(P)$ allows to write $\theta(P_{n,h}) = \theta_0 + \theta_0'(h)/\sqrt{n} + o(1/\sqrt{n})$, in $\mathbb{B}$. By the directional differentiability of ϕ and the Delta-method for directionally differentiable functions,[30]

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

as random elements in $\mathbb{D}$. Next, let l_M be a bounded Lipschitz loss function satisfying Remark 1. By the Portmanteau Theorem,

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

uniformly in h in any finite set $I \subset T(P)$, as $n \to \infty$. Therefore, taking a supremum over such $I \subset T(P)$ yields

$$\begin{aligned}&\sup_{I \subset T(P)} \liminf_{n\to\infty} \sup_{h \in I} \mathbb{E}_{P_{n,h}}\left\{l_M\left(\sqrt{n}(\hat{\phi}_n - \phi(\theta_n(h)))\right)\right\} \\ &\qquad \leqslant \sup_{h \in T(P)} \mathbb{E}\left\{l_M\left(\phi_0'(\mathbb{G}_0 + v_1 + \theta_0'(h)) - \phi_0'(\theta_0'(h)) + v_2\right)\right\}. \qquad \text{(A.14)}\end{aligned}$$

The supremum in the second line of the above display can be equivalently taken over $s \in \theta_0'(T(P))$ and further over the closure of this set in $\mathbb{B}$ (by Lemma A.1), which is

[29] The probability limits under $P_{n,h}$ are the same as under $P_{n,0}$, since $P_{n,0}$ is contiguous with respect to $P_{n,h}$ (Lemma 6.4 in van der Vaart, 2000)

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

equal to $S(\mathbb{G}_0)$. Therefore, provided that the adjustment temrs $\hat{v}_{1,n}, \hat{v}_{2,n}$ converge in probability to the minimizers of the above expression, it follows that

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

$$\leqslant \inf_{(v_1,v_2)\in K} \sup_{s \in S(\mathbb{G}_0)} \mathbb{E}\left\{ l_M \left(\phi_0'(\mathbb{G}_0 + v_1 + s) - \phi_0'(s) + v_2 \right) \right\}. \quad \text{(A.15)}$$

Therefore, the estimator in Equation (A.13) is Locally Asymptotically Minimax.

It remains to show that $\hat{v}_{1,n}, \hat{v}_{2,n}$ defined in the statement of the theorem converge in probability to some minimizers of the LAM risk. In view of Lemma A.9, it suffices to show that Assumptions 1 (identification condition) and 4 (uniform convergence) there are satisfied. Since K is compact and the criterion function is continuous, the identification condition is immediate, so I will show the uniform convergence. Denote:

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

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

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

$$|\hat{g}_n(b, v, s) - g(b, v, s)|$$

$$\leqslant C_M \left(\left|\left| \hat{\phi}'(b + v_1 + s) - \phi_0'(b + v_1 + s) \right|\right| + \left|\left| \hat{\phi}_n'(s) - \phi_0'(s) \right|\right| \right). \quad \text{(A.16)}$$

Let:

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

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

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

$$Q_{4,n}(v) = \sup_{s \in R_n} \mathbb{E}\left(g\left(\mathbb{G}_0, v, s \right) \right),$$

$$Q(v) = \sup_{s \in S(\mathbb{G}_0)} \mathbb{E}\left(g(\mathbb{G}_0, v, s) \right).$$

First, $\sup_{v\in K}|\hat{Q}_{1,n}(v)-\hat{Q}_{2,n}(v)|=o_P(1)$ follows immediately from Lemma A.10 and the fact that $d_H(\hat{R}_n,R_n)=o_P(1)$ by Assumption 5.3. To show that Lemma A.10 can be applied with $f_n(s;v)=\mathbb{E}(\hat{g}_n(\hat{\mathbb{G}}_n^*,v,s)\mid X_1^n)$, note that

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

Second, $\sup_{v\in K}|\hat{Q}_{2,n}(v)-\hat{Q}_{3,n}(v)| = o_P(1)$ follows from the assumed uniform consistency of $\hat{\phi}_n'$ in Assumption 5.2. Indeed, note that Assumption 5.1 implies that $\hat{\mathbb{G}}_n^*$ converges weakly to $\mathbb{G}_0$ unconditionally (see Lemma S.3.1. in the supplemental appendix to Fang and Santos, 2019). Next, fix any $\varepsilon>0$ and $\eta>0$. Since $\mathbb{G}_0$ is tight, there is a compact set $S\subset\mathbb{B}$ such that $P(\mathbb{G}_0\notin S)\leqslant\varepsilon\eta$. Then, by the Portmanteau Theorem, for any $\delta>0$

$$\limsup_{n\to\infty}P(\hat{\mathbb{G}}_n^*\notin S^\delta)\leqslant P(G_0\notin S)\leqslant\varepsilon\eta$$

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

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

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

$$\begin{aligned}&\mathbb{E}\left(\left|\hat{g}_n(\hat{\mathbb{G}}_n^*,v,s)-g(\hat{\mathbb{G}}_n^*,v,s)\right|\,\middle|\,X_1^n\right)\\&\qquad\qquad\leqslant 2M\cdot P(\hat{\mathbb{G}}_n^*\notin S^\delta|X_1^n)+\sup_{b\in S^\delta}|\hat{g}_n(b,v,s)-g(b,v,s)|\end{aligned}\tag{A.17}$$

and, therefore,

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

where $K_n = S + K + R_{l_n,\lambda_n}$. The latter supremum converges in probability to zero by Assumption 5.2.

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

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

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

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

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

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

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

and therefore

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

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

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

for large enough n and some constant C independent of n, where the second inequality follows from the Lipschitz-continuity of $s \mapsto \mathbb{E}(g(\mathbb{G}_0, v, s))$ (Lemma A.7) and Assumption 5.3. By Lemma A.2, a uniformly Lipschitz sequence of functions converging pointwise on a compact set also converges uniformly. Therefore, $\sup_{v\in K}|Q_{4,n}(v) - Q(v)| = o(1)$.

It follows from the preceding discussion that $\sup_{v\in K}|\hat{Q}_n(v) - Q(v)| = o_P(1)$ and, therefore, Lemma A.9 implies that $\vec{d}_H(\hat{V}_n, \mathcal{V}_0) = o_P(1)$, where $\hat{V}_n, \mathcal{V}_0 \subset K$ denote the sets of minimizers of $\hat{Q}_n$ and $Q(v)$ correspondingly within K.[31] Then, for an arbitrary $\hat{v}_n \in \hat{V}_n$, provided that for every subsequence $\hat{v}_{n_k}$ there is a further subsequence $\hat{v}_{n_{k_j}}$ that converges in probability to a constant, the constant must be some $v \in \mathcal{V}_0$. Since (A.15) holds along subsequences $\hat{v}_{n_{k_j}}$, it must hold for the entire sequence as well.

■

Lemma A.9 (Point-wise Consistency of Set Extremum Estimators)**.** *Let $(\mathcal{V}, d)$ be a metric space. Let $\hat{Q}_n(v)$ and $Q(v)$ denote the empirical and population criterion functions, correspondingly. Let $\mathcal{V}_0$ denote the set of maximizers of the population criterion function and $\hat{v}_n$ denote any "almost maximizer" of $\hat{Q}_n$ over a sieve space $\mathcal{V}_{k(n)}$, i.e.*

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

Assume that the following conditions hold.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

■

Lemma A.10 (Replacing The Feasible Set)**.** *Let* $(\mathbb{B}, ||\cdot||_{\mathbb{B}})$ *be a Banach space,* $K \in \mathbb{B}$ *be a compact set and* $f_n : \mathbb{B} \times \mathbb{B} \to \mathbb{R}$ *be a sequence of random functions satisfying, for each* $x_1, x_2 \in \mathbb{B}$*,*

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

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

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

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

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

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

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

■