Adaptive Estimation of Intersection Bounds: a Classification Approach

Vira Semenova*

March 3, 2023

Abstract

This paper studies averages of intersection bounds – the bounds defined by the infimum of a collection of regression functions – and other similar functionals of these bounds, such as averages of saddle values. Examples of such parameters include Frechet-Hoeffding bounds, Makarov (1981) bounds on distributional effects. The proposed estimator classifies covariate values into the regions corresponding to the identity of binding regression function and takes the sample average. The paper shows that the proposed moment function is insensitive to first-order classification mistakes, enabling the use of various regularized/machine learning classifiers in the first (classification) step. The result is generalized to cover bounds on the values of linear programming problems and best linear predictor of intersection bounds.

1 Introduction

Economists are often interested in bounds on parameters when parameters themselves are not point-identified (Manski 1989, 1990), such as quantiles of heterogeneous treatment effects and other measures other than mean. Baseline or pre-treatment covariates often contain helpful information that helps tighten the bounds (Manski and Pepper 2000) on the parameter of interest. The underlying nuisance functions appearing in covariate-specific bounds often create statistical or computational challenges. In this paper, I focus on averages of intersection bounds – the infimum of a collection of nonparametric regression functions – and propose a large sample inference procedure based on a novel moment equation. I show

*Email: vsemenova@berkeley.edu. I am grateful to Bryan Graham, Michael Jansson, Patrick Kline, Lihua Lei, Demian Pouzo, Jim Powell, Rahul Singh, Sophie Sun, Davide Viviano, Christopher Walters and the participants of 2022 California Econometrics Conference for helpful discussions.
that this procedure is first-order insensitive to the misclassification mistake in the identity of the binding constraint.

Let me demonstrate the basic idea using the example Frechet-Hoeffding bounds, a classic example in program evaluation (Manski (1997); Heckman et al. (1997)) literatures. Let $D = 1$ be a binary treatment, let $S(1)$ and $S(0)$ be potential binary outcomes in the treated and control states, and let $S = DS(1) + (1 - D)S(0)$ be the realized outcome. The target parameter is the share of subjects $\Pr(S(1) = S(0) = 1)$ whose outcome is one in both treated and control state (i.e., the always-takers). The sharp upper bound $\pi_U$ on the always-takers’ share is

$$\pi_{at} = \mathbb{E}_X \Pr(S(1) = S(0) = 1 \mid X) \leq \mathbb{E}_X \min_{d \in \{1, 0\}} s(d, X) =: \pi_U$$

$$= \mathbb{E}_X \sum_{d \in \{1, 0\}} \frac{1\{D = d\} \cdot S}{\Pr(D = d \mid X)} 1\{d = \arg \min_{d \in \{1, 0\}} s(d, X)\}$$

where $s(1, x)$ and $s(0, x)$ are the conditional probability of $S = 1$ the treated and control state, respectively. Most of earlier work (Fan and Park (2010), Chernozhukov et al. (2013)) has extensively studied regression approach (1.1). This paper proposes an asymptotic theory based on the moment equation (1.2), arguing that its accommodates a wider set of first-stage plug-in estimators.

The paper focuses on bounds that can be represented as averages of intersection bounds

$$\psi_0 = \mathbb{E}_X \inf_{t \in T} s_0(t, X),$$

where $T$ is a possibly infinite index set and $s_0(t, x)$ are regression functions. The first contribution is to derive the asymptotic theory based on plug-in classifiers

$$t_0^*(x) := \arg\min_{t \in T} s_0(t, x),$$

where the minimizer function $t_0^*(x)$ is estimated from the regression functions $U_{t \in T} s(t, x)$. Assuming the estimates of regression functions $s_0(t, x)$ converge at worst-case $o(N^{-1/4})$ rate, I show that the sharp bound must be first-order insensitive to the classification mistakes. As a result, the proposed estimator is asymptotically equivalent to its oracle counterpart, where the oracle knows the true value of $t_0^*(x)$. Thus, the proposed estimator accommodates regularized procedures that were not previously admissible with the regression approach. Furthermore, this estimator will be the first one (for this parameter) to have its asymptotic variance expressed analytically in closed form (rather than approximated by simulation). This result is the first example in the debiased inference literature where the orthogonality (first-order
insensitivity) property follows from an envelope argument rather than attained by adding a correction term (e.g., Newey (1994)). The result covers Frechet-Hoeffding bounds (Manski (1997); Heckman et al. (1997), Fan and Park (2010)) and Makarov (1981) bounds on distributional effects as special cases.

The leading special case of the bound (1.3) comes from linear programming (LP) problem. A standard form of an LP is

\[ q^\top \beta_0 := \min_{\beta_0 \in \mathbb{R}^p} q^\top \beta_0 \text{ s.to } A\beta_0 = b_0 \]

\[ \beta_0 \geq 0, \]  

where both the LHS matrix \( A \) and the cost parameter \( q \) are known and deterministic while \( b_0 \) is a vector of moments (e.g., population means). If there are baseline covariates available, I propose averaging optimal values \( q^\top \beta_0(x) \) of the conditional LPs

\[ \min_{\beta_0 \in \mathbb{R}^p} q^\top \beta_0 \text{ s.to } A\beta_0 = b_0(x) \]

\[ \beta_0 \geq 0. \]  

and characterize this parameter as a special case of (1.3). Therefore, the proposed bound is weakly tighter than the basic bound (1.5) that does not use covariates. The proposed estimator only requires calculating the vertices of the dual feasible set which does not depend on \( x \). It neither attempts solving the primal (1.6) or dual, nor requires the solution to be available closed form. An immediate application of (1.6) are the Frechet-Hoeffding-type bounds on joint distributions with multi-valued treatments and outcomes, with possibly additional constraints from economic theory expressed in the matrix \( A \).

An important class of LP bounds (e.g., bounds on labor supply responses (Kline and Tartari (2016)) or bounds as in Kamat (2021)) require both the matrix \( A(x) \) and the free vector \( b_0(x) \) to be functions of \( x \), if conditioned on covariates

\[ \min_{\beta_0 \in \mathbb{R}^p} q^\top \beta_0 \text{ s.to } A(x)\beta_0 = b_0(x) \]

\[ \beta_0 \geq 0, \]  

As a result, their target parameters \( \mathbb{E}_X q^\top \beta_0^\ast(X) \) are special cases of expectations of optimal values of Lagrangians

\[ \mathbb{E}_X \sup_{\kappa \in \mathcal{K}} \inf_{t \in T} s_0(t, \kappa, X) = \mathbb{E}_X \inf_{t \in T} \sup_{\kappa \in \mathcal{K}} s_0(t, \kappa, X) = \mathbb{E}_X s_0(t_0(X), \kappa_0(X), X) \]
which moves us outside the main framework (1.3). The proposed estimator is shown to be insensitive to the first-order mistakes in the primal and dual optimizers, extending the oracle property from minimizers $t_0^*(x)$ in (1.4) to saddle points $(t_0^*(x), \kappa_0^*(x))$. The other extensions of (1.3) include best linear predictors of intersection bounds and taking nonlinear transformations of $\inf_{t \in T} s_0(t, x)$, such as trimmed means.

The paper unifies two lines of research that have been previously studied separately. The first one is the literature on bounds, in particular, intersection bounds and bounds on aggregate measures of heterogeneous treatment effects. The second one is the literature on classification, statistical treatment rules and policy learning.

**Bounds, Convex Optimization, and Directionally Differentiable Functionals.** Set identification is a vast area of research, encompassing a wide variety of approaches: linear and quadratic programming, random set theory, support function, and moment inequalities (Manski (1990), Manski and Pepper (2000), Manski and Tamer (2002), Haile and Tamer (2003), Chernozhukov et al. (2007), Beresteanu and Molinari (2008), Molinari (2008), Ciliberto and Tamer (2009), Lee (2009), Stove (2009), Andrews and Shi (2013), Beresteanu et al. (2014), Chandrasekhar et al. (2012), Chernozhukov et al. (2015), Gafarov (2019), Kallus et al. (2020)), see e.g. Molchanov and Molinari (2018) or Molinari (2020) for a review. Most of the results on distributional effects (Makarov (1981), Manski (1997); Heckman et al. (1997), Fan and Park (2010, 2012), Tetenov (2012), Fan et al. (2017), Firpo and Ridder (2019)) focus on identification and/or deriving sharp bounds with covariates, while inference is much less studied. The first discussion of estimators and inference appears in Fan and Park (2010), where, on p.945 they sketch a plug-in of $\Psi_0$ based on moment condition (1.1) without statistical guarantees. Targeting the envelope function $\inf_{t \in T} \mathcal{s}(t, x)$, the work by Chernozhukov et al. (2013) proposes a plug-in approach based on least squares series estimators, where large sample inference is based on the strong approximation of a sequence of series or kernel-based empirical processes. Switching focus from the envelope function to its best linear predictor, Chandrasekhar et al. (2012) proposes a root-$N$ consistent and uniformly asymptotically Gaussian estimator of the target parameter, relying on the first-stage series estimators. Finally, recent work by Lee (2021) focuses on bounds on conditional distributions of treatment effects. That is, most inference work focuses on the envelope function, rather than its mean value, which makes the lack of differentiability of $x \to \min(x, 0)$ at the kink point $x = 0$ a common concern (e.g., Fang and Santos (2018)). Because expectation $\mathbb{E}_X[\cdot]$ in (1.3) is a smoothing operator, this challenge does not apply to the parameter in (1.3). The paper contributes to recently growing literature on bounds coming from linear programming problems (Honoré and Tamer (2006), Andrews et al. (2020), Fang et al. (2020), Dong et al. (2021), Hsieh et al. (2021)). Hsieh et al. (2021) proposes a duality argument similar to the one used here, however,
linking LP to covariates appears to be new. The work by Fang et al. (2020) also uses duality approach for inference / hypothesis testing problem under weaker assumptions that the present manuscript. In contrast, this paper focuses on estimation. Finally, the paper contributes to a growing literature on machine learning for bounds and partially identified models (Kallus and Zhou (2019), Jeong and Namkoong (2020), Bruns-Smith and Zhou (2023), Semenova (2023)) and sensitivity analysis (Dorn and Guo (2021), Dorn et al. (2021), Bonvini and Kennedy (2021), Bonvini et al. (2022)), see e.g., Kennedy (2022) for the review.

**Policy learning and classification.** The Frechet bound (1.1)–(1.2) coincides with the negative first-best welfare in the statistical treatment choice literature (Kitagawa and Tetenov (2018), Mbakop and Tabord-Meehan (2021), Athey and Wager (2021)) up to a constant. However, the minimizer function \( t_0(X) \) and the population bound (1.3) have opposite priorities. The optimal policy (i.e., the classifier \( t_0(X) \)) is the primary target parameter in the treatment choice literature, while this paper focuses on the bound itself. In contrast, most classification and treatment choice papers treat (1.3) as a criterion function. They are primarily interested in the optimal policy \( t_0(X) \) that approximately attains optimal value (1.3), while this value itself is of secondary interest. Likewise, these papers rely on the margin assumption (Mammen and Tsybakov (1999); Tsybakov (2004)) to improve the statistical guarantees of the ERM classifier. Here, I rely on its to control the misclassification bias of the plug-in estimator.

The rest of the paper is organized as follows. Section 2 provides motivating examples and sketches the proposed result. Section 3 formally states the main result. Section 4 presents the extensions, including general case optimization problems, best linear predictor of intersection bounds, and nonlinear smoothing functionals of intersection bounds.

## 2 Set-Up

Many causal parameters of interest can only be bounded from above and below because they are not point-identified. I focus on bounds that can be represented as

\[
\psi_0 := \mathbb{E}_X \inf_{t \in T} s(t, X)
\]

where \( X \) is a covariate vector, \( T \) is a possibly infinite index set, and \( s(t, x) \) is a regression function of \( x \). Examples include Frechet-Hoeffding bounds in (Heckman et al. (1997); Manski (1997)) and Makarov (1981) bounds on distributional effects. The paper’s goal is to develop asymptotically Gaussian inference on \( \psi_0 \) based on the regularized/machine learning classifiers as well as to characterize the asymptotic
variance bound.

**Notation and Definitions.** Let me introduce notation. Suppose each function \( s(t,x) \) is a conditional expectation function of some observed random variable \( g_t(W) \).

\[
s(t,x) = \mathbb{E}[g_t(W) \mid X = x].
\]

The the identity of binding constraint, which is assumed unique a.s., is the minimizer function

\[
t_0(x) := \arg \min_{t \in T} s_0(t,x).
\]

The envelope regression function is

\[
\inf_{t \in T} s_0(t,x) = s_0(t^*_0(x),x).
\]

The envelope moment function is

\[
g(W, \eta) := \sum_{t \in T} g_t(W) 1\{t = t(X)\}, \quad \eta(x) := t(x). \tag{2.2}
\]

Taking conditional expectations of the envelope moment coincides with the envelope regression function:

\[
\mathbb{E}[g(W, \eta_0) \mid X] = \sum_{t \in T} s(t,X) 1\{t = t_0(X)\} = \inf_{t \in T} s_0(t,X),
\]

which implies

\[
\psi_0 = \mathbb{E}_X \mathbb{E}[g(W, \eta_0) \mid X] = \mathbb{E}g(W, \eta_0). \tag{2.3}
\]

The oracle asymptotic variance of \( \psi_0 \)

\[
V_\psi := \mathbb{E} \sum_{t \in T} \mathbb{E}[g_t^2(W) \mid X] 1\{t = t_0(X)\} - \psi_0^2. \tag{2.4}
\]

If the function \( t_0(X) \) was known, the population mean would attain the variance \( V_\psi \).
2.1 Motivating Examples.

Bounds on parameters of joint distributions with fixed marginals. Let $D = 1$ be an indicator for treatment receipt. Let $S(1)$ and $S(0)$ denote the potential binary outcomes if an individual is treated or not, respectively. I assume the standard ignorability assumption.

Assumption 1 (Conditional independence). The vector of potential outcomes is independent of the treatment $D$ conditional on $X$

$$(S(1), S(0), X) \perp \! \! \! \perp D \mid X$$

and the propensity score

$$\mu_1(X) = \Pr(D = 1 \mid X)$$ (2.5)

is known. Example 2.1 describes Frechet-Hoeffding (Manski (1997); Heckman et al. (1997)) bounds on the always-takers’ share.

Example 2.1. Frechet-Hoeffding bounds Suppose Assumption [1] holds with a binary outcome $S \in \{1, 0\}$. Then, the conditional always-takers’ share

$$\pi_{at}(x) := \Pr(S(1) = S(0) = 1 \mid X = x)$$

is bounded as

$$\pi_{\min}(x) := \max(s(0, x) + s(1, x) - 1, 0) \leq \pi_{at}(x) \leq \min(s(0, x), s(1, x)) =: \pi_{\max}(x)$$ (2.6)

Aggregating the bound over covariate space gives the sharp lower and upper bounds

$$\pi_L := \mathbb{E}\max(s(0, X) + s(1, X) - 1, 0) \leq \Pr(S(1) = S(0) = 1) \leq \mathbb{E}\min(s(0, X), s(1, X)) =: \pi_U$$ (2.7)

Thus, $\pi_U$ is a special case of (2.1) with the index set

$$T = \{1, 0\}$$

and $s(t, x) = \Pr(S = t \mid X = x)$ for $t \in T$. An envelope moment equation for $\pi_U$ is given by

$$g(W, \eta_0) := \sum_{t \in \{1, 0\}} \frac{D = t}{\Pr(D = t \mid X)} S1\{t = \eta_0(X)\}$$ (2.8)
The asymptotic variance in (2.4) reduces to
\[
V_\psi = E \left[ \sum_{t \in \{0,1\}} \frac{s(t,X)}{\mu_t(X)} 1 \{t_0(X) = t\} \right] - \pi^2_U, \tag{2.9}
\]
When the propensity score is
\[
\mu_1(x) = \mu_0(x) = 1/2, \tag{2.10}
\]
the asymptotic variance reduces to a function of the bound itself \(V_\psi = \pi_U(2 - \pi_U)\). If the strong overlap condition
\[
0 < \kappa \leq \inf_{x \in X} \mu_1(x) \leq \sup_{x \in X} \mu_1(x) \leq 1 - \kappa < 1. \tag{2.11}
\]
holds, Assumption 3.3 is automatically satisfied. If the other assumptions hold, the statement of Theorem 3.1 holds with \(\psi_0 = \pi_U\) in (2.7) and \(g(W, \eta_0)\) in (2.8) and \(V_\psi\) in (2.9).

**Example 2.2.** **Makarov (1981) bounds on distributional effects** Consider the setup of Example 2.1 with \(S(1)\) and \(S(0)\) being continuously distributed outcomes. Let \(F_1(\cdot \mid x)\) and \(F_0(\cdot \mid x)\) be a conditional CDF of \(S \mid D = 1, X = x\) and \(S \mid D = 0, X = x\), respectively. Let \(F_{S(1)-S(0)}(t)\) be the CDF of the treatment effect \(S(1) - S(0)\), and let \(F_{S(1)-S(0)}(t \mid x)\) be its conditional analog. Makarov (1981) shows that the sharp upper bound on \(F_{S(1)-S(0)}(t)\) is given by
\[
\psi_U := 1 + E \inf_{t \in T} (F_1(t \mid X) - F_0(t \mid X), 0) \tag{2.12}
\]
The sharp upper bound \(\psi_U\) is a special case of \(\psi_0\) in (2.1) with \(T = R \cup \{\alpha\}\) and
\[
s(t,x) := E[1 \{S \leq t\} \mid D = 1, X = x] - E[1 \{S \leq t\} \mid D = 0, X = x], \quad t \in R,
\]
and
\[
s(\alpha,x) := E[0 \mid X = x] = 0.
\]
The envelope moment equation for the upper bound \(\psi_U\) is
\[
g(W, \eta) := \sum_{t \in T} \left( \frac{D1 \{S \leq t(X)\}}{\mu_1(X)} - \frac{(1-D)1 \{S \leq t(X)\}}{\mu_0(X)} \right) 1 \{t(x) = \arg\min_{t \in T} s(t,x)\} \tag{2.13}
\]
The asymptotic variance is
\[
V_\psi = \mathbb{E} \sum_{t \in T \neq \{0\}} \left( \frac{F_1(t(X))}{\mu_1(X)} + \frac{F_0(t(X))}{\mu_0(X)} \right) 1\{t(X) = t\} - \psi^2.
\]
(2.14)

Suppose the strong overlap condition (2.11) holds. Under Assumptions 3.1 and 3.2, the statement of Theorem 3.1 holds with \(\psi_0 = \psi_U\) in (2.12) and \(g(W, \eta_0)\) in (2.13) and \(V_\psi\) in (2.14).

**Example 2.3. Linear Programming (LP) in a special case** Consider an LP (linear programming) problem in the standard form:
\[
\beta^q_0(x) := \min_\beta - q' \beta
\]
subject to
\[
A \beta = b_0(x),
\beta \geq 0,
\]
where \(q \in \mathbb{R}^k\) and \(A \in \mathbb{R}^{k \times p}\) are known deterministic quantities that do not change with \(x\). The RHS vector \(b_0(x)\) is an expectation function
\[
b_0(x) = \mathbb{E}[S \mid X = x].
\]
The target parameter is
\[
\psi_0 = \mathbb{E} q' \beta^q_0(X).
\]
(2.16)

I show that the target parameter (2.16) is a special case of (2.1). Consider the dual form of the problem (2.15)
\[
\max_{\nu, \lambda} -b^T_0(x) \nu \text{ subject to } \begin{pmatrix} A^T & -I_p \end{pmatrix} (\nu; \lambda)' - q = 0,
\lambda \geq 0.
\]
(2.17)

By strong duality, the primal and dual optimal values are equal
\[
-b^T_0(x) \nu^* (x) = q' \beta^q_0 (x).
\]
(2.18)

Notice that the dual feasible set defined by (2.17) does not depend on \(x\). This set is the polytope with
at most \(\binom{p+k}{p}\) vertices, where each vertex corresponds to a Basic Feasible Solution (BFS). Assuming the primal LP has a finite optimal solution for each \(x\), the dual LP must have achieve its optimal value in one of the vertices. Taking the index set \(T := \mathcal{T}(A; q)\) as the vertex set and regression function as

\[ s(t, x) = b_0^T(x) v_t, \quad (v_t; \lambda_t) \in \mathcal{T}(A; q) =: T \]  

(2.19)
gives

\[ b_0^T(x) v_0^* (x) = \inf_{(v_t; \lambda_t) \in \mathcal{T}(A; q)} b_0^T(x) \cdot v_t =: \inf t \in T s(t, x), \]

and the target parameter is

\[ \psi_0 = \mathbb{E} q' \beta_0^T(X) = \mathbb{E} \inf t \in T s(t, X). \]  

(2.20)

A formal proof of (2.20) is given in Lemma 1.4. The orthogonal moment for \(\psi_0\) takes the form

\[ g(W, \eta_0) = \sum_{t \in \mathcal{T}(A; q)} S^T \nu_t \{ s(t, x) = \arg \min_{t \in T} s(t, x) \} \]  

(2.21)

The asymptotic variance \(V_{\psi}\) reduces to

\[ V_{\psi} = \mathbb{E} \sum_{t \in T} \mathbb{E}[(S^T \nu_t)^2 \mid X \} \{ s(t, x) = \arg \min_{t \in T} s(t, x) \} \mid X] - \psi_0^2. \]  

(2.22)

**Remark 2.1.** Example 2.1 is a special case of Example 2.3 with

\[ q = (1, 0, 0, 0), \quad A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \]

the RHS as the expectation vector-function

\[ b_0(x) = \mathbb{E}[S \mid X = x], \quad S := \left( \frac{D \cdot S}{\mu_1(X)}, \frac{(1-D) \cdot S}{\mu_0(X)}, 1 \right). \]

The set of basic feasible solutions (i.e., vertices of dual feasible set) is

\[ \mathcal{T}(A; q) = \{ (v_1; \lambda_1), (v_2; \lambda_2) \} = \left\{ (1, 0, 0, 0, 1, 0, 0)', (0, 1, 0, 0, 0, 1, 0)' \right\}. \]
where the first 3 coordinates correspond to the shadow prices of each of the 3 constraint. The vector \( \nu^*_1(x) = (1, 0, 0) \) corresponds to \( x : s(1,x) < s(0,x) \), where solution \( \min(s(1,x), s(0,x)) = s(1,x) \) is attained at the corner solution with the first row of \( A \) being active constraint. Likewise, \( \nu^*_2(x) = (0, 1, 0)' \) if \( \min(s(1,x), s(0,x)) = s(0,x) \) and the second row of \( A \) is active constraint.

### 2.2 Other Examples

A large number of LP problems do not fall into the framework (2.15). In these problems, conditioning on covariates results in both LHS matrix \( A(x) \) and the RHS vector \( b_0(x) \) being \( x \)-dependent. The target parameters take the form

\[
\begin{align*}
\min_{\beta} &- q' \beta \\
\text{s.t.} & A(x)\beta = b_0(x), \\
& \beta \geq 0,
\end{align*}
\]

where the matrix \( A(x) \) depends on \( x \). I study this parameter as a special case of generic optimization problem Example 2.4.

#### Example 2.4. Generic optimization problem.

Consider a standard form optimization problem:

\[
\begin{align*}
\min f_0(\beta) \text{ s.t. } f(\beta, x) &\leq 0, \quad h(\beta, x) = 0,
\end{align*}
\]

where \( \beta \in \mathbb{R}^p \). The constraint functions are \( f(\beta, x) = (f_1(\beta, x), \ldots, f_{mf}(\beta, x)) \in \mathbb{R}^{mf} \) and \( h(\beta, x) = (h_{mf+1}(\beta, x), h_{mf+2}(\beta, x), \ldots, h_m(\beta, x)) \in \mathbb{R}^{m_f} = \mathbb{R}^m \). The target parameter takes the form

\[
\psi_0 = E_X f_0(\beta^*_0(X)).
\]

In what follows, assume there exist an observed random variable \( S_j(\beta) \) such that

\[
\begin{align*}
f_j(x, \beta) &= E[S_j(\beta) \mid X = x], \quad j = 1, 2, \ldots, mf \\
h_j(x, \beta) &= E[S_j(\beta) \mid X = x], \quad j = mf + 1, \ldots, m.
\end{align*}
\]

The target parameter is

\[
\psi_0 = E_X f_0(\beta^*_0(X)).
\]
For each \( x \), consider the Lagrangian function. I represent it as a regression function

\[
s_0(\kappa, t, x) = f_0(\beta) + \sum_{j=1}^{m_f} \lambda_j s_j(\beta) + \sum_{j=m_f+1}^{m} \nu_j h_j(\beta, x),
\]

(2.27)

where \( t = \beta \) is the inner minimization argument and \( \kappa = (\nu, \lambda) \) is the outer maximization argument, respectively. The dual function is

\[
g(\kappa, x) = \inf_{t \in T} s_0(\kappa, t, x).
\]

(2.28)

The dual maximization problem is

\[
\max_{\kappa} g(\kappa, x) \text{ s.t. } \lambda \geq 0.
\]

(2.29)

Assuming strong duality holds (e.g., Slater condition holds) for each \( x \), the primal and dual objectives coincide

\[
g(\kappa_0(x), x) = f_0(\beta_0(x)) = f_0(t_0(x)) \quad \forall x \in \mathcal{X},
\]

(2.30)

which implies that \((t_0(x), \kappa_0(x))\) is a saddle point of the regression function (2.27). As a result, the (2.26) reduces to

\[
\psi_0 = \mathbb{E}_X \sup_{\kappa \in \mathcal{K}} \inf_{t \in T} s_0(\kappa, t, X) = \mathbb{E}_X \inf_{\kappa \in \mathcal{K}} \sup_{t \in T} s_0(\kappa, t, X).
\]

(2.31)

For each \((\kappa, t)\), define the moment function

\[
g_{\kappa, t}(W) := f_0(\beta) + \sum_{j=1}^{m_f} \lambda_j s_j(\beta) + \sum_{j=m_f+1}^{m} \nu_j h_j(\beta, x),
\]

\( t := \beta, \kappa := (\nu, \lambda). \)

The proposed moment function is

\[
\psi_0 = \mathbb{E} \sum_{t \in T} \sum_{\kappa \in \mathcal{K}} 1\{\kappa = \kappa_0(X), t = t_0(X)\} g_{\kappa, t}(W),
\]

where \((\kappa_0(x), t_0(x))\) is the saddle point of \(s_0(\kappa, t, x)\). The asymptotic variance is

\[
V_\psi := \mathbb{E} \sum_{t \in T} \sum_{\kappa \in \mathcal{K}} 1\{\kappa = \kappa_0(X), t = t_0(X)\} [g_{\kappa, t}^2(W) \mid X] - \psi_0^2.
\]

(2.32)
Note that the max-min mapping \( x \rightarrow \max_{\kappa \in \mathcal{K}} \min_{t \in T} s_0(\kappa, t, x) \) is neither convex nor concave in \( x \). As a result, the sharp bounds may not correspond to aggregating over full covariate space. That said, there could be other motivation to condition on covariates, for example, to sustain the unconfoundedness assumption.

**Example 2.5. Best Linear Predictor of Intersection bounds** Suppose the envelope function is a smooth function of \( x \) that can be approximated as

\[
\inf_{t \in T} s(t, X) = p(X)' \beta_0 + R(X),
\]

where \( p(X) \) is a \( d \)-vector of basis functions, \( \inf_{t \in T} s(t, X) \) is the envelope function, \( R(X) \) is the approximation error and \( \beta_0 \) is the best linear predictor. The target parameter \( \beta_0 \) is the best linear predictor

\[
\beta_0 = (\mathbb{E} p(X) p(X)')^{-1} \mathbb{E} p(X) \inf_{t \in T} s_0(t, X)
\]

\[
= (\mathbb{E} p(X) p(X)')^{-1} \mathbb{E} p(X) g(W, \eta_0),
\]

where \( g(W, \eta_0) \) is in (2.8). The pointwise and uniform asymptotic theory for the best linear predictor based on (2.34) is discussed in Theorem 4.2.

**Example 2.6. Bounds on \( \mathbb{E}(S(1) - S(0))^+ \)** Consider the setup of Example 2.2. Tetenov (2012) and Fan et al. (2017) study sharp lower bound on the partially identified parameter \( \mathbb{E}(S(1) - S(0))^+ \). The bound takes the form

\[
\mathbb{E}_X \int_R (F_0(t \mid X) - F_1(t \mid X))^+ dt = \mathbb{E}_X \int_R \max(F_0(t \mid X) - F_1(t \mid X), 0) dt,
\]

where \( F_1(t \mid x) \) and \( F_0(t \mid x) \) are the CDFs of \( S \mid D = 1, X = x \) and \( S \mid D = 0, X = x \), respectively. Fubini theorem gives

\[
\mathbb{E}_X \int_R \max(F_0(t \mid X) - F_1(t \mid X), 0) dt = \int_R \int_X \max(F_0(t \mid x) - F_1(t \mid x), 0) f_X(x) dx dt,
\]

which suggests an orthogonal moment

\[
g(W, \eta_0) =: \int_R g_t(W, \eta_0) dt,
\]
where
\[ g_t(W, \eta_0) = \left( \frac{(1-D)1\{S \leq t\}}{\mu_0(X)} - \frac{D1\{S \leq t\}}{\mu_1(X)} \right) 1\{F_0(t \mid X) - F_1(t \mid X) > 0\} \]

**Example 2.7. Interval-Valued Outcome**

Semenova (2023) studies parameters that are linear in an unobserved scalar outcome \( Y \). The identified set takes the form
\[ B = \{ \beta = \Sigma^{-1}E[V(\eta_0)Y, \ Y_L \leq Y \leq Y_U] \}, \quad (2.35) \]

where the random vector \( V(\eta_0) \in \mathbb{R}^d \) depends on a nuisance function \( \eta_0 \) and the matrix \( \Sigma \) is identified.

The target parameter is the support function \( \sigma(q) := \sup_{b \in B} q'b = \sup_{Y: Y_L \leq Y \leq Y_U} q'\Sigma^{-1}E[V(\eta_0)Y = q'\Sigma^{-1}E[V(\eta_0)Y^*(\eta_0, q)], \quad (2.36) \]

where \( Y^*(\eta_0, q) \) is the selection (a particular element of the random set \([Y_L, Y_U]\)) that maximizes (2.36).

Beresteanu and Molinari (2008); Bontemps et al. (2012) have shown that \( Y^*(\eta_0, q) \) has a closed-form expression
\[ Y^*(\eta, q) = \begin{cases} Y_L & q'\Sigma^{-1}V(\eta) < 0 \\ Y_U & q'\Sigma^{-1}V(\eta) > 0 \end{cases} \]

Envelope theorem gives
\[ \partial_\eta \sigma(q) = q'\Sigma^{-1}E[\partial_\eta V(\eta_0)Y^*(\eta_0, q)], \]

where the selection \( Y^*(\eta_0, q) \) is insensitive to the classification error. Let \( X \) denote the argument of the nuisance function \( \eta_0 \). If \( V(\eta) \) is a random variable given \( X = x \), the equivalence of margin assumption (3.1) is the bound on the conditional density of \( V(\eta) \mid X = x \), and the equivalent of \( s_N^* \) is the mean square rate \( s_N \) (Partially Linear Model with Interval-Valued Outcome, Semenova (2023)). If \( V(\eta) \) and \( \eta \) are functions of the same argument (e.g. Average Partial Derivative, Kaido (2017)), asymptotic theory is based on the combination of margin assumption and the worst-case \( s_N^* \) rate.

### 2.3 Envelope theorem. Zero derivative property.

In this section, I consider the plug-in classifiers, that is, the classifiers based on the estimated regression functions \( \{s(t, x)\}_{t \in T} \). The plug-in estimator of \( t(X) \):
\[ \eta(x) := t(x) := \arg \inf_{t \in T} s(t, x), \]
where \( s(t,x) \) are estimates of \( s_0(t,x) \). Then, the conditional misclassification effect on the true function \( s_0(t,x) \) is

\[
\tau_0(x) := s_0(\eta(x), x) - s_0(t_0(x), x).
\]

(2.37)

A covariate value \( x \) is misclassified if and only if \( \tau_0(x) > 0 \).

Let me introduce a two-stage estimator of \( \psi_0 \). In the first stage, I estimate the regression functions and construct the plug-in classifier. In the second stage, I compute the sample estimate of the moment function.

Definition 1 (Cross-Fitting).

1. For a random sample of size \( N \), denote a \( K \)-fold random partition of the sample indices \( [N] = \{1, 2, \ldots, N\} \) by \( (J_k)_{k=1}^K \), where \( K \) is the number of partitions and the sample size of each fold is \( n = N/K \). For each \( k \in [K] = \{1, 2, \ldots, K\} \) define \( J_k^c = \{1, 2, \ldots, N\} \setminus J_k \).

2. For each \( k \in [K] \), construct an estimator \( \hat{\eta}_k = \hat{\eta}(V_{i \in J_k^c}) \) of the nuisance parameter \( \eta_0 \) using only the data \( \{V_j : j \in J_k^c\} \). Take \( \hat{\eta}(X_i) := \hat{\eta}_k(X_i), \quad i \in J_k \).

Definition 2 (Plug-in estimator of \( \psi_0 \)). Given the fitted values of the estimated classifier \( (\hat{\eta}(X_i))_{i=1}^N \), define

\[
\hat{\psi}(\hat{\eta}) := N^{-1} \sum_{i=1}^N g(W_i, \hat{\eta}) := N^{-1} \sum_{i=1}^N \sum_{t \in T} g_t(W_i) 1\{t(X_i) = \hat{\eta}(X_i)\}.
\]

(2.38)

Under the conditions discussed below, the plug-in estimator \( \hat{\psi}(\hat{\eta}) \) is equivalent to its oracle version

\[
\sqrt{N} (N^{-1} \sum_{i=1}^N g(W_i, \hat{\eta}) - N^{-1} \sum_{i=1}^N g(W_i, \eta_0)) = o_P(1),
\]

(2.39)

where the oracle knows the true value of first-stage arg min function. Therefore, \( \hat{\psi}(\hat{\eta}) \) is asymptotically Gaussian

\[
N^{-1/2} \sum_{i=1}^N g(W_i, \eta_0) - \psi_0 \Rightarrow N(0, V_\psi).
\]

where \( V_\psi \) in (2.4) coincides with the oracle asymptotic variance.

Remark 2.2 (Envelope property). The oracle property (2.39) can be interpreted as an application of envelope theorem-type argument. Consider Example 2.1. Define the conditional average treatment effect (CATE)

\[
\zeta(x) = s(1,x) - s(0,x).
\]

(2.40)
Consider a class of sets $G \subseteq X$ determined by plug-in estimates of conditional ATE

$$G(\zeta) := \{x : \zeta(x) \geq 0\}.$$  

Then, the Frechet-Hoeffding upper bound can be expressed as

$$\pi_U := \min_{G(\zeta)} \pi_U(G) = \pi_U(G^*(\zeta_0)),$$

where

$$\pi_U(G) = \mathbb{E} \left[ \frac{1-D}{\mu_0(X)} 1 \{ X \in G \} + \frac{D}{\mu_1(X)} 1 \{ X \notin G \} \right].$$

The optimal value of the bound $\pi_U = \pi_U(G^*)$ is first-order insensitive with respect to perturbations in $\zeta$.

### 3 Main Result

#### 3.1 Assumptions

**ASSUMPTION 3.1** (Bound on the first-stage $\ell_\infty$ rate). There exists a sequence $s^*_N = o(N^{-1/4})$ such that the following worst-case rate bound holds.

$$\sup_{s(t,x) \in S_t} \sup_{t \in T} \sup_{x \in X} |s(t,x) - s_0(t,x)| \leq s^*_N = o(N^{-1/4}).$$

Assumption 3.1 requires the first-stage estimator $s(t,x)$ of $s_0(t,x)$ converges at $o(N^{-1/4})$ rate. When stated in $L_2$ rate, the $o(N^{-1/4})$ rate is a classic assumption in the semiparametric literature. Examples of estimators obeying $\ell_\infty$ rate bound include $\ell_1$-regularized estimators in Belloni et al. (2017).

**ASSUMPTION 3.2** (Margin assumption). Assume that there exist finite positive constants $\bar{B}_f, \delta \in (0, \infty)$ such that

$$\Pr\left( \min_{t \in T} \min_{x} s_0(t,X) - \min_{t \in T} s_0(t,X) \leq t \right) \leq \bar{B}_f t, \quad \forall t \in (0, \delta). \quad (3.1)$$

Assumption 3.2 states the generalization of margin assumption (Mammen and Tsybakov (1999); Tsybakov (2004)) for a possibly infinite index set, as proposed in Qian and Murphy (2011). While it does allow the maximizer of $s(t,x)$ to consist of more than 1 element, it restricts the difference in conditional mean functions between the envelope function $s(t(x),x)$ and the best suboptimal function at $x$
\[
\min_{t \in T} \arg\min_{x} s(t,x). \quad \text{In the binary case with } T = \{1, 0\}, \text{ the margin condition (3.1) reduces to }
\]
\[
\Pr(0 < |s(1,X) - s(0,X)| \leq t) \leq \bar{B} t, \quad \forall t \in (0, \delta).
\] (3.2)

For example, (3.2) holds if \( s(1,X) - s(0,X) \) is continuously distributed with a bounded density.

Define the conditional second moment
\[
\rho_0(t,x) = \mathbb{E}[g^2(W) \mid X = x], \quad t \in T.
\] (3.3)

**ASSUMPTION 3.3** (Bounded Second Moment). There exists a constant \( 0 < \bar{B} < \infty \) such that
\[
\sup_{t \in T} \sup_{x \in \mathcal{X}} \rho_0(t,x) \leq \bar{B}.
\] (3.4)

Assumption 3.3 is a standard regularity condition, ensuring that the conditional second moment is bounded uniformly over \( T \) and \( \mathcal{X} \).

**Theorem 3.1** (Asymptotic Theory for \( \psi_0 \) in (2.1)). Under Assumptions 3.1–3.3 the envelope orthogonal moment obeys the oracle property (2.39). As a result,
\[
\sqrt{N} \left( N^{-1} \sum_{i=1}^{N} g(W_i, \hat{\eta}) - \psi_0 \right) \Rightarrow_d N(0, V_{\psi}),
\]
where the oracle asymptotic variance \( V_{\psi} \) in (2.4).

Theorem 3.1 is my main result. It states the asymptotically Gaussian approximation for the sharp bound parameter. It also provides an analytic expression for the asymptotic variance that has not previously made available.

Consider the asymptotic variance \( V_{\psi} \) in (2.4). Unless \( V_{\psi} \) is a function of \( \psi_0 \) (e.g., (2.10) holds), the asymptotic variance \( V_{\psi} \) is first-order sensitive to the misclassification error. The sample analog estimator
\[
\widehat{V}_{\psi} := N^{-1} \sum_{i=1}^{N} \rho(t,X_i) 1\{t = \arg\min_{s(t,X)}\} - \hat{\psi}^2,
\] (3.5)

(3.5) is first-order sensitive to the estimation error in \( s(t,x) \). Lemma 3.2 discusses its convergence rate.

**Lemma 3.2** (Consistency and Rate of the Variance Estimator). Assuming \( s_0(t,x) \) and \( \rho_0(t,x) \) are estimated at \( s_N \) and \( \rho_N \), respectively, under the conditions of Theorem 3.1
\[
|\widehat{V}_{\psi} - V_{\psi}| = O_P(\rho_N + s_N + N^{-1/2}).
\]
3.2 Linear Programming (special case).

Algorithm 1 describes the estimator of $\psi_0$ in Example 2.3.

**Algorithm 1** Upper bound on parameters determined by LP with constant LHS matrix $A(x) = A$.

Input: direction $q$, matrix $A$, cross-fitted values $(\hat{b}(X_i))_{i=1}^N$.

1: Calculate the set of Basic Feasible Solutions $\mathcal{T}(A; q) = \{(v, \lambda)\}$ for the dual LP whose constraints are

$$(A^\top - I_p) (v; \lambda) - q = 0.$$  

2: Estimate the identity of binding constraint

$$\hat{\eta}(X_i) := \min_{(v; \lambda) \in \mathcal{T}(A; q)} \hat{b}(X_i) v_i$$

3: Report: the sample estimate

$$\hat{\psi} := N^{-1} \sum_{i=1}^N S^\top \hat{\eta}(X_i).$$ (3.6)

**ASSUMPTION 3.4** (LP regularity conditions). (1) The estimator $b(x)$ of $b_0(x)$ converges uniformly in $\ell_2$ norm, that is

$$\sup_{b \in B_N} \sup_{x \in X} \|b(x) - b_0(x)\| = o(N^{-1/4}).$$ (3.7)

(2) The linear combinations of vector $b_0(X)$ obey margin assumption, that is,

$$\Pr(\inf_{v \in \mathbb{R}^p, ||v||=1} |v^\top b_0(X)| \leq t) \leq \bar{B}t$$ (3.8)

(3) The variance matrix of $S$ is bounded in the operator norm, that is

$$\sup_{x \in X} \max \text{eig} \mathbb{E}(SS^\top | X = x) \leq \bar{B} \text{ a.s. in } x.$$ (3.9)

(4) The matrix $A \in \mathbb{R}^{k \times p}$ has full row rank with $k$. (5) The problem (2.15) is feasible with finite optimal value $\beta_q(x)$ for each covariate value.

**Corollary 3.1** (Linear Programming Bounds). Suppose Assumption 3.4 holds. Then, the Theorem 3.1 holds for the $\hat{\psi}$ of (3.6) and $V_{\hat{\psi}}$ in (2.22).
4 Extensions

In this section, I generalize my baseline result. Section 4.1 extends the theory of plug-in estimators from minimizers to saddle-points, covering Examples 2.4 as a special case. Section 4.2 generalizes main result to best linear predictor of intersection bounds.

4.1 Generic optimization problem

Consider Example 2.4. The target parameter in (2.31) is

$$\psi_0 = \mathbb{E}_X \max_{\kappa \in \mathcal{K}} \min_{t \in \mathcal{T}} s_0(t, \kappa, X).$$

A point \((\kappa_0(x), t_0(x))\) is a saddle point of \(s_0(t, \kappa, x)\) if it obeys

$$s_0(t_0(x), \kappa, x) \leq s_0(t_0(x), \kappa_0(x), x) \leq s_0(t, \kappa_0(x), x) \quad \forall \kappa \in \mathcal{K} \forall t \in \mathcal{T}.$$

Given a regression estimate \(s(t, \kappa, x)\) of \(s_0(t, \kappa, x)\), let \((\hat{\kappa}(x), \hat{t}(x))\) be the saddle point of estimated regression function \(s(t, \kappa, x)\). Given the cross-fit estimated saddle points \((\hat{\kappa}(X_i), \hat{t}(X_i))_{i=1}^N\), I propose the following plug-in estimator of \(\psi_0\) in (2.31).

Definition 3 (Plug-in estimator of \(\psi_0\) for generic LP problems). Given the fitted values \((\hat{\kappa}(X_i), \hat{t}(X_i))_{i=1}^N\), define

$$\hat{\psi} := N^{-1} \sum_{i=1}^N \sum_{\kappa \in \mathcal{K}} \sum_{t \in \mathcal{T}} g_{\kappa,t}(W_i) 1 \{ \kappa = \hat{\kappa}(X_i), t = \hat{t}(X_i) \}. \quad (4.1)$$

**Assumption 4.5** (Bound on the first-stage rate). There exists a sequence \(s_N^\infty = o(N^{-1/4})\) such that the following worst-case rate bound holds.

$$\sup_{s(t, x) \in \mathcal{S}_x} \sup_{\kappa \in \mathcal{K}} \sup_{t \in \mathcal{T}} \sup_{x \in \mathcal{X}} |s(t, \kappa, x) - s_0(t, \kappa, x)| \leq s_N^\infty = o(N^{-1/4}).$$

**Assumption 4.6** (2d Margin Assumption). Assume that there exist finite positive constants \(\bar{B}_f, \delta \in \mathbb{R}^d\).
Suppose \( \min_{\kappa \in K, t \neq t_0} |s_0(t, \kappa, X) - s_0(t_0, \kappa, X)| \leq B_f t \) \( \forall t \in (0, \delta) \). \hspace{1cm} (4.2)

Suppose \( \min_{t \in T, \kappa \neq \kappa_0} |s_0(t, \kappa, X) - s_0(t, \kappa_0, X)| \leq B_f t \) \( \forall t \in (0, \delta) \). \hspace{1cm} (4.3)

Theorem 4.1 (Asymptotic Theory for General LP Problems). Suppose Assumptions 4.5–4.6 hold and for some \( \bar{B} \in (0, \infty) \)

\[
\sup_{t \in T} \sup_{\kappa \in K} \sup_{x \in X} \mathbb{E}[g^2(W_t, \hat{\eta}) | X = x] \leq \bar{B}.
\]

Then, the estimator is asymptotically Gaussian

\[
\sqrt{N}(N^{-1} \sum_{i=1}^{N} g(W_i, \hat{\eta}) - \psi_0) \Rightarrow N(0, V_\psi),
\]

where the oracle asymptotic variance \( V_\psi \) in (2.32).

Theorem 4.1 is my second main result. It establishes the asymptotic theory for the expectation of the optimal values of a generic optimization problem beyond those representable in (2.1). It extends the argument of Theorem 3.1 from minimizers to saddle points.

4.2 Best Linear Predictor of Intersection bounds.

Consider Example 2.5. The target parameter is the best linear predictor of intersection bounds

**Definition 4 (Best Linear Predictor).** Given the first-stage fitted values \( (\hat{\eta}(X_i))_{i=1}^{N} \), define

\[
\hat{\beta} := \left( \frac{1}{N} \sum_{i=1}^{N} p(X_i)p(X_i)^t \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} p(X_i)g(W_i, \hat{\eta}(X_i)).
\] \hspace{1cm} (4.4)

**Assumption 4.7 (Regularity conditions on the basis functions and first-stage rate).** (1) The sup-norm of the basis functions \( \xi_d := \sup_{x \in \mathcal{X}} \|p(x)\| = \sup_{x \in \mathcal{X}} (\sum_{j=1}^{d} p_j(x)^2)^{1/2} \) grows sufficiently slow:

\[
\sqrt{\frac{\xi_d^2 \log N}{N}} = o(1), \hspace{1cm} d \xi_d^2 (\log N)^{1/2} = o(N^{-1/2})
\]

(2) There exists a sequence of finite constants \( l_d, r_d \) such that the norms of the misspecification error are controlled as follows:

\[
\|r_g\|_{p_2} := \sqrt{\int r_g(x)^2 dP(x)} \lesssim r_d \quad \text{and} \quad \|r_g\|_{p_\infty} := \sup_{x \in \mathcal{X}} |r_g(x)| \lesssim l_d r_d, \hspace{1cm} (\xi_d^2 \log N/N)^{1/2} \cdot (1 + l_d r_d \sqrt{d}) = o(1)
\]
There eigenvalues of $Q = \mathbb{E} p(X) p(X)^\top$ are bounded away from above and below.

**Theorem 4.2** (Pointwise and Uniform Inference on Intersection bounds). Under Assumptions 3.2 and 4.7, the proposed estimator of best linear predictor is asymptotically normal:

$$
\lim_{N \to \infty} \sup_{t \in \mathbb{R}} \left| \Pr \left( \frac{\sqrt{N} \alpha'(\hat{\beta} - \beta_0)}{\sqrt{\alpha' \Omega \alpha}} < t \right) - \Phi(t) \right| = 0. 
$$

Moreover, if the approximation error is negligible relative to the estimation error, namely $\sqrt{N} r_\pi(x_0) = o(\|\Omega^{1/2} p(x_0)\|)$, then $\hat{g}(x) = p(x_0)' \hat{\beta}$ is asymptotically normal:

$$
\lim_{N \to \infty} \sup_{t \in \mathbb{R}} \left| \Pr \left( \frac{\sqrt{N} (\hat{g}(x_0) - g(x_0))}{\sqrt{p(x_0)' \Omega p(x_0)}} < t \right) - \Phi(t) \right| = 0. 
$$

Theorem 4.2 extends the basic result of Theorem 3.1 to accommodate linear smoothers of intersection bounds, such as best linear predictor. It follows from Theorem 3.1 of Semenova and Chernozhukov (2021).

**A Appendix A. Proofs**

**Lemma 1.3.** The equality (2.4) holds.

**Proof of Lemma 1.3.** The result follows from the law of total variance

$$
\text{Var}(g(W, \eta_0)) = \mathbb{E} \left[ \text{Var}(g(W, \eta_0) \mid X) \right] + \text{Var} \left( \inf_{t \in \mathcal{I}} s_{t_0}(t, X) \right)
\quad
= \mathbb{E} \left[ \sum_{t \in \mathcal{I}} \mathbb{E} \left[ S_{t_0}(X) \mid X \right] \mathbb{I}\{t = t_0(X)\} \right] - \mathbb{E} \left[ \sum_{t \in \mathcal{I}} \mathbb{I}\{t = t_0(X)\} s_{t_0}(t, X) + \psi_0 - \psi_0^2 \right]_

\right]
\quad
\blacksquare

**Lemma 1.4 (Characterization of $\psi_0$ in Example 2.3).** Suppose the following conditions hold. (1) $A \in \mathbb{R}^{k \times p}$ has full row rank with $k$. (2) The problem (2.15) is feasible with finite optimal value $\beta_0^q(x)$ a.s. in $x$.

**Proof.** (1) Consider the dual problem (2.17) for any point $x$. For an LP with linear constraints, strong duality reduces to feasibility, which is assumed in (2). Therefore, strong duality (2.18) holds, and the dual Lagrange problem must be feasible with a finite optimal value for each $x$.

(2) Notice that the dual feasible set does not depend on $x$. Let $\mathcal{J}(A; q)$ be the set of its BFS (basic feasible solutions). The cardinality of the vertex set $|\mathcal{J}(A; q)| \leq \binom{p+q}{p}$. By (1), the dual program has a
finite optimal value, and, therefore, it must have an optimal BFS. Therefore,

\[ b_T^0(x) v_T^0(x) = \inf_{(v_t, x_t) \in T(A, q)} b_T^0(x) v_t =: \inf_{t \in T} s(t, x), \]

where \( s(t, x) \) are as in (2.19).

Proof of Theorem 3.1. Step 1. Define

\[ t(X) = \arg \min s(t, X) \]
\[ t_0(X) = \arg \min s_0(t, X) \]
\[ \tau(X) := s(t(X), X) - s(t_0(X), X) \leq 0 \]
\[ \tau_0(X) := s_0(t(X), X) - s_0(t_0(X), X) \geq 0 \]

Suppose \( \tau(X) \neq 0 \). Then,

\[ \tau(X) < 0 \leq \tau_0(X) \Rightarrow \tau(X) - \tau_0(X) < -\tau_0(X) \leq 0. \]

Assumption 3.2 gives

\[ \Pr(0 \leq \tau_0(X) \leq t) \leq \Pr(s_0(t(X), X) - s_0(t_0(X), X) \leq t) \]
\[ \leq \Pr \left( \min_{t \in \arg \min_{s_0(t, X)}} s_0(t, X) - s_0(t_0(X), X) \leq t \right) \leq \bar{B} f t. \]

Note that

\[ \mathbb{E}[g(W; \eta) - g_0(W) | X] = s_0(t(X), X) - s_0(t_0(X), X) = \tau_0(X). \]

Therefore, the bias is bounded as

\[ |\mathbb{E}[g(W; \eta) - g_0(W)]| \leq \mathbb{E}|\tau(X)|1\{\tau(X) - \tau_0(X) \leq -\tau_0(X) \leq 0\} \]
\[ \leq \mathbb{E}|\tau(X) - \tau_0(X)|1\{\tau(X) - \tau_0(X) \leq -\tau_0(X) \leq 0\} \]
\[ \leq 2\lambda^\infty \Pr(\tau(X) - \tau_0(X) \leq -\tau_0(X) \leq 0) \]
\[ \leq 2\lambda^\infty \Pr(-2\lambda^\infty \leq -\tau_0(X) \leq 0) \leq \bar{B} f ((\lambda^\infty)^2), \]
where Assumption 3.1 gives
\[ \tau(X) - \tau_0(X) < -\tau_0(X) \leq 0 \Rightarrow -2\delta_N^0 < -\tau_0(X) \leq 0. \]

The variance is bounded as
\[
\mathbb{E}[(g(W; \eta) - g_0(W))^2 \mid X] \leq 2(\mathbb{E}[S^2_{t(X)} \mid X] + \mathbb{E}[S^2_{\tau(X)} \mid X])1\{\tau(X) - \tau_0(X) \leq -\tau_0(X) \leq 0\}
\]
\[
\leq \bar{C} \Pr(\tau(X) - \tau_0(X) \leq -\tau_0(X) \leq 0) = \bar{C}\delta_N^0.
\]

\[\square\]

**Proof of Corollary 3.1.** The set \( \mathcal{T}(A; q) \) is a finite set. The constants
\[ \max_{(v_1, \lambda_1) \in \mathcal{T}(A; q)} \|v_1\| := \bar{C} < \infty \]
and
\[ \min_{(v_1, \lambda_1), (v_2, \lambda_2), v_1 \neq v_2 \in \mathcal{T}(A; q)} \|v_1 - v_2\| := c > 0 \]
are bounded away from above and below. Invoking the bound
\[ \sup_{t} \sup_{x \in \mathcal{X}} |s(t, x) - s_0(t, x)| \leq \max_{(v_1, \lambda_1) \in \mathcal{T}(A; q)} \|v_1\| \sup_{x \in \mathcal{X}} \|b(x) - b_0(x)\| \leq \bar{C}\|b(x) - b_0(x)\| \]
(1.1)
and Assumption 3.4 (1) verifies Assumption 3.1. Next, Assumption 3.2 reduces to
\[
\Pr\left( \min_{t \in T \backslash \arg \min_{t \in T} s_0(t, X)} s_0(t, X) - \min_{t \in T} s_0(t, X) \leq t \right)
\]
\[
=^a \Pr\left( \min_{(v_1, \lambda_1), (v_2, \lambda_2) \in \mathcal{T}(A; q); v_1 \neq v_2} \| (v_1 - v_2)'b_0(X) \| \leq t \right)
\]
\[
\leq^b \Pr\left( \inf_{v \in \mathbb{R}^p, \|v\| = 1} \| v' b_0(X) \| \leq t/c \right) \leq \tilde{B}/\sqrt{q},
\]
where (a) follows from the definition of set \( T = \{ v_1, (v_1, \lambda_1) \in \mathcal{T}(A; q) \} \) (b) follows from the inequality
\[
\min_{(v_1, \lambda_1), (v_2, \lambda_2) \in \mathcal{T}(A; q); v_1 \neq v_2} \| (v_1 - v_2)'b_0(X) \|
\]
\[
\geq \inf_{v \in \mathbb{R}^p, \|v\| = 1} \| v' b_0(X) \| \min_{(v_1, \lambda_1), (v_2, \lambda_2) \in \mathcal{T}(A; q); v_1 \neq v_2} \| v_1 - v_2 \| = \inf_{v \in \mathbb{R}^p, \|v\| = 1} \| v' b_0(X) \| c
\]

23
and (c) follows from Assumption 3.4(2). Finally, Assumption 3.3 follows from Assumption 3.4(3):

\[
\sup_{t \in T} \sup_{x \in \mathcal{X}} \mathbb{E}[\mathcal{G}_t^2(W) \mid X = x] \leq \sup_{(v; \lambda)} \mathbb{E}[(S^t v)_t]^2 \mid X = x] \\
\leq \sup_{(v; \lambda) \in \mathcal{J}(\mathcal{A}), q} v_t' \mathbb{E}[SS_t \mid X = x] v_t \\
\leq \sup_{(v; \lambda) \in \mathcal{J}(\mathcal{A}), q} \|v_t\|^2 B \leq C^2 B.
\]

\[\blacksquare\]

**Proof of Theorem 4.1**  Step 1. By strong duality assumed in (2.18), the primal and dual optimal points \((t_0(x), \kappa_0(x))\) form a saddle point of the true regression function \(s_0(t, \kappa, x)\). Therefore,

\[s_0(\kappa, t_0(x), x) \leq s_0(\kappa_0(x), t_0(x), x) \leq s_0(\kappa_0(x), t, x) \quad \forall \kappa \forall x. \quad (1.2)\]

Likewise, \((\kappa(x), t(x))\) is a saddle point of \(s(t, \kappa, x)\). Therefore,

\[s(\kappa_0(x), t(x), x) \leq s(\kappa(x), t(x), x) \leq s(\kappa(x), t_0(x), x). \quad (1.3)\]

By Assumption 4.5 there exists \(N\) large enough such that

\[\sup_{\kappa \in \mathcal{K}} \sup_{t \in T} \sup_{x \in \mathcal{X}} |s(\kappa, t, x) - s_0(\kappa, t, x)| \leq s_N^0 \quad (1.4)\]

for \(s_N^0 = O(N^{-1/4})\). Combining the definitions of saddle-points \((1.2) - (1.3)\) gives

\[|s(\kappa(x), t(x), x) - s_0(\kappa_0(x), t_0(x), x)| \leq s_N^0. \quad (1.5)\]

Invoking \((1.5)\) and \((1.4)\) gives

\[|s_0(\kappa(x), t(x), x) - s_0(\kappa_0(x), t_0(x), x)| \leq |s(\kappa(x), t(x), x) - s_0(\kappa(x), t(x), x)| \\
+ |s(\kappa(x), t(x), x) - s_0(\kappa_0(x), t_0(x), x)| \leq 2s_N^0.\]

**Step 2.** In this step, I bound the probability of misclassification event. Define

\[\tau_0(x) = s_0(\kappa(x), t(x), x) - s_0(\kappa_0(x), t_0(x), x).\]
A covariate value \( x \) is misclassified if and only if \( \tau_0(x) \neq 0 \). Consider a covariate value \( x : \tau_0(x) > 0 \).

Decompose

\[
\tau_0(x) = (s_0(\kappa(x), t(x), x) - s_0(\kappa(x), t_0(x), x)) + s_0(\kappa(x), t_0(x), x) - s_0(\kappa_0(x), t_0(x), x) > 0
\]

\[
\leq 0
\]

where the second term \( \tau_0^-(x) \leq 0 \) by saddle point property (1.2) of \((\kappa_0(x), t_0(x))\). If \( \tau_0(x) > 0 \), it must be that \( \xi^+(x) > 0 \). Invoking the saddle point property (1.3) gives

\[
s(\kappa(x), t(x), x) - s(\kappa(x), t_0(x), x) \geq 0,
\]

which implies

\[
0 < \xi^+(x)
\]

\[
\leq s_0(\kappa(x), t(x), x) - s_0(\kappa(x), t_0(x), x) + s(\kappa(x), t(x), x) - s(\kappa(x), t_0(x), x)
\]

\[
\leq |s(\kappa(x), t(x), x) - s_0(\kappa(x), t(x), x)| + |s(\kappa(x), t_0(x), x) - s_0(\kappa(x), t_0(x), x)| \leq 2\delta^0_N.
\]

Assumption 4.6 gives

\[
Pr(0 < \xi^+(x) < t) \leq Pr(\min_{\kappa \in \mathcal{X}} \min_{t \neq t_0} |s_0(\kappa, t, X) - s_0(\kappa, t_0, X)| \leq t) \leq \bar{B}t,
\]

which implies \( Pr(0 < \xi^+(x) < 2\delta^0_N) = O(\delta^0_N) \). Collecting the terms gives

\[
\mathbb{E}\tau_0(X)I\{\tau_0(X) < 0\} \leq E\xi^+(X)I\{\tau_0(X) < 0\}
\]

\[
\leq E(|s(\kappa(x), t(X), X) - s_0(\kappa(x), t(X), X)| + |s(\kappa(x), t_0(X), X) - s_0(\kappa(x), t_0(X), X)|)I\{\tau_0(X) > 0\}
\]

\[
\leq 2\delta^0_N Pr(\tau_0(X) > 0) \leq 2\delta^0_N Pr(\xi^+(X) > 0) = O(r^2_N).
\]

**Step 3.** Likewise, if \( \tau_0(x) < 0 \), we decompose

\[
\tau_0(x) = (s_0(\kappa(x), t(x), x) - s_0(\kappa_0(x), t(x), x)) + s_0(\kappa_0(x), t(x), x) - s_0(\kappa_0(x), t_0(x), x) < 0
\]

\[
\geq 0
\]

\[
= \xi^-(x) + \tau_0^+(x),
\]

which implies \( \xi^-(x) < 0 \). The rest of the argument follows similarly to Step 2, except for the Assumption
is invoked as

$$
\Pr(0 < \zeta^{-}(X) < t) \leq \Pr(\min_{i \in T} \min_{\kappa \neq \kappa_0} \lvert s_0(\kappa, t, X) - s_0(\kappa_0, t, X) \rvert \leq t) \leq \bar{B}t,
$$

which implies \( \Pr(0 < \zeta^{-}(X) < s_N^\infty) = O(s_N^\infty) \).

References


Kamat, V. (2021). Identifying the effects of a program offer with an application to head start.


