Minimax Instrumental Variable Regression and L_2 Convergence Guarantees without Identification or Closedness

Andrew Bennett AWB222@CORNELL.EDU
Cornell University

Nathan Kallus KALLUS@CORNELL.EDU
Cornell University

Xiaojie Mao MAOXJ@SEM.TSINGHUA.EDU.CN
Tsinghua University

Whitney Newey WNEWEY@MIT.EDU
Massachusetts Institute of Technology

Vasilis Syrgkanis VSYRGK@STANFORD.EDU
Stanford University

Masatoshi Uehara MU223@CORNELL.EDU
Cornell University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

In this paper, we study nonparametric estimation of instrumental variable (IV) regressions. Recently, many flexible machine learning methods have been developed for instrumental variable estimation. However, these methods have at least one of the following limitations: (1) restricting the IV regression to be uniquely identified; (2) only obtaining estimation error rates in terms of pseudometrics (*e.g.,* projected norm) rather than valid metrics (*e.g.,* L_2 norm); or (3) imposing the so-called closedness condition that requires a certain conditional expectation operator to be sufficiently smooth. In this paper, we present the first method and analysis that can avoid all three limitations, while still permitting general function approximation. Specifically, we propose a new penalized minimax estimator that can converge to a fixed IV solution even when there are multiple solutions, and we derive a strong L_2 error rate for our estimator under lax conditions. Notably, this guarantee only needs a widely-used source condition and realizability assumptions, but not the so-called closedness condition. We argue that the source condition and the closedness condition are inherently conflicting, so relaxing the latter significantly improves upon the existing literature that requires both conditions. Our estimator can achieve this improvement because it builds on a novel formulation of the IV estimation problem as a constrained optimization problem.

Keywords: Instrumental variables, Inverse problems, Empirical risk minimization

1. Introduction

Instrumental variable (IV) estimation is an important problem in many applications. Examples include causal inference (Angrist and Imbens, 1995; Newey and Powell, 2003; Deaner, 2018; Cui et al., 2020), missing data problems (Wang et al., 2014; Miao et al., 2015), asset pricing models (Chen et al., 2014; Christensen, 2017; Escanciano et al., 2020), dynamic discrete choice models (Kalouptsidi et al., 2021), and reinforcement learning (Liao et al., 2021; Uehara et al., 2021).

In this paper, we focus on the estimation of nonparametric IV (NPIV) regression (Newey and Powell, 2003). This problem involves three sets of variables X, Y, and Z that take values in compact

Euclidean sets D_X, D_Y, and D_Z, respectively. In the original IV estimation problem, X stands for endogenous variables, Y stands for an outcome variable, and Z stands for exogenous IVs. We define $L_2(X), L_2(Z)$ as the L_2 spaces of functions of X, Z respectively, defined in terms of their distributions. We are interested in solving the following equation with respect to $h \in L_2(Z)$:

$$\mathbb{E}\left[Y - h(X) \mid Z\right] = 0.$$

This equation can be alternatively written as $\mathcal{T}h = r_0$, where $r_0(Z) = \mathbb{E}[Y \mid Z]$, and $\mathcal{T} : L_2(X) \to L_2(Z)$ is a bounded linear operator that maps every $h \in L_2(X)$ to $\mathbb{E}\left[h(X) \mid Z\right] \in L_2(Z)$. Here both the function r_0 and the operator $\mathcal{T}$ are unknown. Instead, we only have access to a set of independent and identically distributed observations $\mathcal{D} \coloneqq \{X_i, Y_i, Z_i\}_{i=1}^n$.

There has been a surge in interest in NPIV regressions. A number of classical works have proposed sieve or kernel-based estimators (e.g., Carrasco et al., 2007; Horowitz, 2011; Newey, 2013; Newey and Powell, 2003; Chen, 2007). However, NPIV estimation is notoriously difficult because it is an ill-posed inverse problem. In particular, the solution to the NPIV equation $\mathcal{T}h = r_0$ may not be unique, and even if it is unique, the solution may depend on the underlying data distribution discontinuously (Carrasco et al., 2007). Therefore, existing works typically assume that the NPIV solution is unique (Andrews, 2017; Newey and Powell, 2003). Even if it is not the case, they restrict the linear operator $\mathcal{T}$ and the NPIV solution (Florens et al., 2011; Chen, 2021). A widely used restriction is the source condition, which assumes that the IV solution belongs to a subspace defined by the operator $\mathcal{T}$ (e.g., Carrasco et al., 2007; Cavalier, 2011; Chen and Reiss, 2011). Under these conditions, the estimators proposed in these classic literature can have strong theoretical guarantees. However, these traditional nonparametric estimators do not allow for the integration of modern, flexible general function approximation methods such as neural networks or tree-based methods.

To overcome this limitation, recent works have proposed various algorithms that can accommodate general function approximation. These algorithms typically employ two function classes, $\mathcal{H}$ and $\mathcal{G}$. In particular, the function class $\mathcal{H}$ is the hypothesis class for the solution to the NPIV equation $\mathcal{T}h = r_0$. The function class $\mathcal{G}$, often referred to as a witness function class or discriminator class, is introduced to witness how much each given function h violates the NPIV equation. Then, NPIV estimators are defined as solutions to a minimax optimization problem (Lewis and Syrgkanis, 2018; Bennett et al., 2019; Dikkala et al., 2020; Liao et al., 2020; Muandet et al., 2020):

$$\arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g)$$

where $L(h, g)$ is an objective function mapping from $\mathcal{H} \times \mathcal{G}$ to $\mathbb{R}$.

Although highly flexible, these minimax approaches have several limitations. First, they typically assume that the solution to the NPIV equation $\mathcal{T}h = r_0$ is unique. However, this assumption can be easily violated if the instrumental variables are not very strong (Andrews and Stock, 2005; Andrews et al., 2019), and they usually do not hold in proximal causal inference (Kallus et al., 2021). Secondly, the minimax estimators may not give strong L_2 error rate guarantees, and instead only have error rate guarantees in terms of a weaker projected mean squared error (MSE) (Dikkala et al., 2020). However, even when the projected MSE vanishes to zero, the minimax estimator may not converge to any fixed IV solution since the projected MSE is a pseudometric unlike the L_2 metric. Thirdly, current minimax estimators typically need some form of closedness condition, such as $\mathcal{T}h \in \mathcal{G}$ for any $h \in \mathcal{H}$ (Dikkala et al., 2020; Liao et al., 2020) or other close variant (Bennett et al., 2022). However, this assumption may impose stringent restrictions on the operator $\mathcal{T}$, noting that $\mathcal{G}$ must be

Table 1: Summary of current literature of minimax estimation with general function approximation. Our goal is to solve $\mathcal{T}h = r_0$ with respect to h with unknown $\mathcal{T}$ and r_0. We denote its set of solutions by $\mathcal{H}_0$ and the least norm solution by h_0. Estimators are defined as solutions to certain minimax optimizations $\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} L(h, g)$ where $\mathcal{H}$ and $\mathcal{G}$ are hypothesis classes. For simplicity, we focus on comparison for VC classes $\mathcal{H}$ and $\mathcal{G}$ (while the results in both our and these papers deal with general function classes) and for source condition with exponent 1. We let $\mathcal{T}^*$ be the adjoint of $\mathcal{T}$, $\bar{\mathcal{G}}_0 = \{\bar{g}_0 : \mathcal{T}^*\bar{g}_0 = h_0\}$ (nonempty under source condition), $h_{0,\alpha} = \arg\min_h \|\mathcal{T}h - r_0\|_2^2 + \alpha\|h\|_2^2$, and $\alpha_0 > 0$ a positive number. Note our condition is strictly weaker than that of Bennett et al. (2022).

	Primary assumptions	Guarantee	Rate
Dikkala et al. (2020)	realizability $\mathcal{H}_0 \cap \mathcal{H} \neq \emptyset$, closedness $\mathcal{T}\mathcal{H} \subset \mathcal{G} + r_0$	Projected MSEs	$n^{-1/2}$
Liao et al. (2020)	source $h_0 \in \mathcal{R}(\mathcal{T}^*\mathcal{T})$, uniqueness of h_0, realizability $h_{0,\alpha} \in \mathcal{H}\ \forall \alpha \leq \alpha_0$, closedness $\mathcal{T}\mathcal{H} \subset \mathcal{G} + r_0$	L_2 rates	$n^{-1/6}$
Bennett et al. (2022)	source $h_0 \in \mathcal{R}(\mathcal{T}^*\mathcal{T})$, realizability $h_0 \in \mathcal{H}$, $\bar{\mathcal{G}}_0 \cap \mathcal{G} \neq \emptyset$ and closedness $\mathcal{T}^*\mathcal{G} \subset \mathcal{H}$	L_2 rates	$n^{-1/4}$
This work	source $h_0 \in \mathcal{R}(\mathcal{T}^*\mathcal{T})$ realizability $h_0 \in \mathcal{H}$, $\bar{\mathcal{G}}_0 \cap \mathcal{G} \neq \emptyset$	L_2 rates	$n^{-1/4}$

a restricted class to ensure bounded statistical complexity. In particular, the closedness assumption is at odds with the widely used source condition, since we will show that the closedness assumption is more plausible when the spectrum of $\mathcal{T}$ decays more slowly while the source condition is more plausible when the spectrum decays more rapidly.

To the best of our knowledge, all current approaches incorporating general function approximation for IV problems suffer from at least one of the three limitations listed above. In this paper, we propose the first method that avoids all three of these limitations. Specifically, we do not assume that the NPIV solution is unique, and instead we target the least norm solution h_0. This is a standard approach for inverse problems with non-unique solutions (Florens et al., 2011; Babii and Florens, 2017; Chen, 2021; Bennett et al., 2022). We show that our proposed estimator can converge to the least norm IV solution and derive its L_2 error rate guarantee. These theoretical guarantees only need the fairly standard source condition and realizability assumptions (*i.e.*, well-specification of $\mathcal{H}$ and $\mathcal{G}$). Table 1 summarizes the assumptions and guarantees in our paper and related ones.

Our proposed estimator and its theory are grounded in the novel insight that finding the least norm solution h_0 to $\mathcal{T}h = r_0$ can be viewed as a constrained optimization problem. In particular, we show that the least norm solution can be uniquely identified as a saddle point of the minimax optimization of the Lagrangian. Although previous minimax estimators also leverage minimax optimization, their inner maximization is used to approximate the projected MSE $\mathbb{E}[([\mathcal{T}h](Z) - r_0(Z))^2]$, which necessitates the closedness assumption. In contrast, the inner maximization in our methods results from the method of Lagrange multipliers, and it does not need the closedness assumption. Interestingly, we prove that the source condition is the sufficient and necessary condition for the existence of stationary Lagrange multipliers and thus the saddle point to our minimax optimization problem. This also reveals a new role of the source condition widely used in inverse problems.

Our paper is organized as follows. In Section 2, we present our setup of IV estimation and the limitations of current works in this setting. In Section 3, we introduce our minimax estimator by

framing the problem as a constrained optimization problem. In Section 4, we demonstrate that the minimax optimization identifies the least norm solution given infinite data. In Section 5, we present the finite-sample error guarantee, i.e., L_2 convergence rate. In Section 6, we compare our estimator and theory to those in closely related works. Finally, we conclude our paper in Section 7.

1.1. Related Works

Instrumental variable estimation has received considerable attention as a subclass of inverse problems, as detailed in the works of Carrasco et al. (2007); Cavalier (2011); Newey (2013); Ito and Jin (2014).

Even when the operator $\mathcal{T}$ and response r_0 are known, nonparametric instrumental variable estimation poses significant difficulties due to its ill-posed nature. The ill-posedness often refers to the presence of one or more of the following characteristics: (1) the absence of solutions, (2) the existence of multiple solutions, and (3) the discontinuity of the inverse of $\mathcal{T}$. To address these challenges, various regularization techniques have been proposed, such as compactness of the solution space (Newey and Powell, 2003), Tikhonov regularization, and Landweber–Fridman regularization (Carrasco et al., 2007; Cavalier, 2011). In practical settings where $\mathcal{T}$ and r_0 are unknown, a range of estimators have been proposed in the literature, including series-based estimators (Ai and Chen, 2003; Hall and Horowitz, 2005; Blundell et al., 2007; Chen and Reiss, 2011; Darolles et al., 2011; Chen and Pouzo, 2012; Florens et al., 2011; Chen, 2021), kernel-based estimators (Hall and Horowitz, 2005; Horowitz, 2007), and RKHS-based estimators (Singh et al., 2019; Muandet et al., 2020).

Recently, there has been growing interest in the application of general function approximation techniques, such as deep neural networks and random forests, to instrumental variable problems in a unified manner (Dikkala et al., 2020; Lewis and Syrgkanis, 2018; Bennett et al., 2019; Zhang et al., 2020). Among these approaches, Dikkala et al. (2020); Liao et al. (2020); Bennett et al. (2022) provide finite-sample convergence rate guarantees. Specifically, Liao et al. (2020) establishes L_2 convergence by linking minimax optimization with Tikhonov regularization under the assumption of the source condition. Bennett et al. (2022) establishes an L_2 convergence guarantee under the source condition from a distinct perspective. Notably, the assumptions we need are strictly weaker than those of Bennett et al. (2022). Dikkala et al. (2020) guarantees convergence in terms of projected mean squared error without the source condition; however, this guarantee is insufficient to identify a specific element when the solution is not unique. These works (Dikkala et al., 2020; Liao et al., 2020; Bennett et al., 2022) rely on the so-called closedness assumption, which imposes restrictions on the smoothness of the operator $\mathcal{T}$ via the witness class. This assumption has been the subject of considerable discussion in the context of offline reinforcement learning, with researchers exploring ways to relax it (Chen and Jiang, 2019; Uehara et al., 2020; Foster et al., 2021; Huang and Jiang, 2022). In this paper, we examine the relaxation of this assumption in a more general IV setting. This is of importance since the source condition and closedness are inherently conflicting.

We note that there are a number of alternative approaches for integrating machine learning into instrumental variable estimation (Hartford et al., 2017; Yu et al., 2018; Xu et al., 2020; Liu et al., 2020; Kato et al., 2021; Lu et al., 2021). However, to the best of our knowledge, these approaches do not offer an L_2 convergence rate guarantee in the absence of the assumption of uniqueness.

2. Problem Setup

We aim to solve the following equation with respect to h:

$$\mathcal{T}h = r_0 \tag{1}$$

where $r_0(Z) = \mathbb{E}[Y \mid Z] \in L_2(Z)$ is an unknown function and $\mathcal{T} : L_2(X) \to L_2(Z)$ is an unknown conditional expectation operator that maps any $h \in L_2(X)$ to $\mathbb{E}[h(X) \mid Z]$. Note that $\mathcal{T}$ is a bounded operator since its norm is upper-bounded by 1 via Jensen's inequality. Moreover, we use $\mathcal{T}^\star : L_2(Z) \to L_2(X)$ to denote the adjoint operator of $\mathcal{T}$, *i.e.*, $\langle g, \mathcal{T}h\rangle_{L_2(Z)} = \langle \mathcal{T}^\star g, h\rangle_{L_2(X)}$ for any $h \in L_2(X), g \in L_2(Z)$ where $\langle\cdot,\cdot\rangle_{L_2(X)}$ and $\langle\cdot,\cdot\rangle_{L_2(Z)}$ are inner products over $L_2(X)$ and $L_2(Z)$, respectively. It is known that $\mathcal{T}^*$ is given by $[\mathcal{T}^* g](X) = \mathbb{E}[g(Z) \mid X]$ for any $g \in L_2(Z)$ (Carrasco et al., 2007). Importantly, here we do not assume compactness of $\mathcal{T}$, because compactness is violated whenever X, Z include common variables, as is the case in many applications (Deaner, 2018; Cui et al., 2020). Moreover, we denote the range space of $\mathcal{T}$ by $\mathcal{R}(\mathcal{T})$, i.e., $\mathcal{R}(\mathcal{T}) = \{\mathcal{T}h : h \in L_2(X)\}$.

Throughout this work, we assume that there exists a solution to Equation (1).

Assumption 1 (Existence of solutions)**.** *We have $r_0 \in \mathcal{R}(\mathcal{T})$, i.e., $\mathcal{N}_{r_0}(\mathcal{T}) := \{h \in \mathcal{H} : \mathcal{T}h = r_0\} \neq \emptyset$.*

Most of the existing literature further assumes that $\mathcal{T}$ is injective and the solution to Equation (1) is unique. However, even in this case, Equation (1) still corresponds to an ill-posed inverse problem, since the inverse operator $\mathcal{T}^{-1}$ is generally unbounded, so the NIPV solution can be very sensitive to even slight perturbations to the data distributions. Without further restrictions, we can only obtain an estimator $\hat{h}$ with convergence guarantee in terms of the projected MSE $\mathbb{E}[\{\mathcal{T}\hat{h} - r_0\}^2(Z)] = \mathbb{E}[\{\mathcal{T}(\hat{h} - h)\}^2(Z)]$ for $h \in \mathcal{N}_{r_0}(\mathcal{T})$. However, the projected MSE is only a pseudometric. Hence, even if $\mathbb{E}[\{\mathcal{T}(\hat{h} - h)\}^2(Z)]$ vanishes to zero, the estimator $\hat{h}$ may not converge to a fixed point. Furthermore, the projected MSE is weaker than the valid metric such as the L_2 metric. Indeed, according to Jensen's inequality, we have $\mathbb{E}[\{\hat{h} - h)\}^2(X)] \geq \mathbb{E}[\{\mathcal{T}(\hat{h} - h)\}^2(Z)]$. However, the other direction generally does not hold. Thus $\mathbb{E}[\{\hat{h} - h)\}^2(X)]$ may not vanish even when $\mathbb{E}[\{\mathcal{T}(\hat{h} - h)\}^2(Z)]$ does.

In many problems, L_2 rate guarantees are preferable or even necessary (Hall and Horowitz, 2005; Chen and Reiss, 2011; Kallus et al., 2021; Uehara et al., 2021). In order to achieve L_2 convergence, we need to further restrict the ill-posedness of the NPIV problem. One common way is to restrict the magnitude of the ill-posedness measure $\sup_{h \in \mathcal{H}} \frac{\mathbb{E}[\{h-h')\}^2(X)]}{\mathbb{E}[\{\mathcal{T}(h-h')\}^2(Z)]}$ for any solution $h' \in \mathcal{N}_{r_0}(\mathcal{T})$, where $\mathcal{H}$ is the function class used to obtain the estimator (*e.g.*, Dikkala et al., 2020; Chen and Pouzo, 2012). This allows us to translate projected MSE guarantees to corresponding L_2 error rates under the uniqueness of Equation (1).

However, in this paper, we do not assume a unique solution to Equation (1), because it may not hold in many practical settings. In particular, uniqueness is violated when instrumental variables are weak (Andrews and Stock, 2005; Andrews et al., 2019). For instance, when the spaces D_X and D_Z are discrete and the cardinality of D_Z exceeds that of D_X, uniqueness generally does not hold. Moreover, uniqueness is usually violated in proximal causal inference, as Kallus et al. (2021) demonstrates in various examples. When solutions are non-unique, Equation (1) becomes even more ill-posed. In this case, existing estimators may still have projected MSE guarantees, but obtaining L_2 rate guarantees becomes much more difficult. In particular, the ill-posedness measure is generally infinity and thus uninformative. Most of the existing estimators do not necessarily converge to any particular solution in $\mathcal{N}_{r_0}(\mathcal{T})$ in terms of the L_2 metric.

Given that there may be (infinitely) many solutions in $\mathcal{N}_{r_0}(\mathcal{T})$, we propose to target a particular solution that achieves the least norm, that is,

$$h_0 = \underset{h \in \mathcal{N}_{r_0}(\mathcal{T})}{\arg\min}\ 0.5\langle h, h\rangle_{L_2(X)}. \tag{2}$$

This least norm solution is well-defined as it is the projection of the origin in $L_2(X)$ onto a closed affine space $\mathcal{N}_{r_0}(\mathcal{T}) \subseteq L_2(X)$. We formalize this in the following lemma.

Lemma 1. *Suppose Assumption 1 holds. Then the least norm solution $h_0 \in \mathcal{N}_{r_0}(\mathcal{T})$ uniquely exists, and $\{h_0\} = \overline{\mathcal{R}(\mathcal{T}^\star)} \cap \mathcal{N}_{r_0}(\mathcal{T})$, where $\overline{\mathcal{R}(\mathcal{T}^\star)}$ is the closure of the range space $\mathcal{R}(\mathcal{T}^\star)$.*

We note that some of the existing literature also targets the least norm solution when the IV equation admits non-unique solutions (Florens et al., 2011; Santos, 2011; Chen, 2021), but they all focus on classic sieve or kernel-based estimators. The only exception is Bennett et al. (2022) as they employ general function approximation while allowing for non-unique solutions. But as we discuss in Section 6.3, their method requires a closedness assumption that puts strong restrictions on the operator $\mathcal{T}$. In this paper, we propose a new estimator for the least norm solution h_0 with a strong L_2 convergence guarantee. Importantly, our estimator accommodates general function approximation but does not need the closedness assumption, thereby improving upon the existing literature.

3. Penalized Minimax Instrumental Variable Regression

In this section, we propose our estimator for the least norm solution h_0 in Equation (2). To this end, we first provide a reformulation of the solution h_0. Note that

$$h_0 = \underset{h \in L_2(X)}{\arg\min}\, 0.5\langle h, h\rangle_{L_2(X)},\ \text{subject to } \mathcal{T}h = r_0.$$

This is a constrained optimization problem over the Hilbert space $L_2(X)$. Following the method of Lagrange multipliers, we can consider an alternative minimax optimization:

$$h_0 = \underset{h \in L_2(X)}{\arg\min} \sup_{g \in L_2(Z)} L(h, g), \quad L(h, g) \coloneqq 0.5\langle h, h\rangle_{L_2(X)} + \langle r_0 - \mathcal{T}h, g\rangle_{L_2(Z)}, \tag{3}$$

where g corresponds to a Lagrange multiplier.

In Equation (3), the objective function $L(h, q)$ is unknown since the two inner products involve the unknown function r_0, the unknown operator $\mathcal{T}$, and the unknown distribution of X and Z. To construct an estimator based on Equation (3), we first rewrite the inner products into expectations with respect to the distribution of X and Z:

$$\langle h, h\rangle_{L_2(X)} = \mathbb{E}\left[h^2(X)\right], \ \ \langle r_0 - \mathcal{T}h, g\rangle_{L_2(Z)} = \mathbb{E}\left[(Y - h(X))\, g(Z)\right].$$

Then we can replace the unknown expectations with empirical averages, and restrict the functions h and g to some classes $\mathcal{H} \subset [D_X \to \mathbb{R}]$ and $\mathcal{G} \subset [D_Z \to \mathbb{R}]$. This leads to the following estimator:

$$\hat{h}_{\text{mn}} \in \underset{h \in \mathcal{H}}{\arg\min} \max_{g \in \mathcal{G}} L_n(h, g), \quad L_n(h, g) \coloneqq 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n\left[(Y - h(X))\, g(Z)\right], \tag{4}$$

where $\mathbb{E}_n[\cdot]$ stands for the empirical average operator based on sample data $\mathcal{D} = \{X_i, Y_i, Z_i\}$. For example, we have $\mathbb{E}_n[h^2(X)] = \frac{1}{n}\sum_{i=1}^n h^2(X_i)$. Notably, the term $\mathbb{E}_n[h^2(X)]$ in Equation (4) can

be viewed as a penalization term, so we call our estimator a penalized minimax estimator. The role of this penalization term is later discussed in Theorem 1.

The estimator $\hat{h}_{\text{mn}}$ in Equation (4) has a minimax optimization formulation. The computational perspective will be discussed in Section C. This is in line with many recent machine learning IV estimators with general function approximation (see a review in Section 1.1). However, our minimax optimization in Equation (4) is motivated by the method of Lagrange multipliers, while existing minimax estimators are based on fundamentally different principles. As a result, our objective function $L_n(h, g)$ differs from those used in existing minimax estimators. In particular, our minimax estimator requires quite different conditions, as we will discuss in Section 6.

To justify the objective function in (4), we need to further guarantee that

$$h_0 = \arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g). \tag{5}$$

In Section 4, we establish Equation (5) under fairly mild conditions. Based on this, we then further derive the L_2 convergence rate of our proposed estimator $\hat{h}_{\text{mn}}$.

4. Identification of the Least Norm Solution

In this section, we establish that our proposed minimax formulation can indeed identify the least norm solution h_0 as shown in Equation (5). We start with introducing a key assumption for our result, and then present our identification result under this assumption.

4.1. Source Condition

Our identification crucially depends on the following source condition.

Assumption 2 (Source condition)**.** *The function r_0 satisfies that $r_0 \in \mathcal{R}(\mathcal{T}\mathcal{T}^\star)$.*

Assumption 2 further strengthens Assumption 1 in that it restricts r_0 to a smaller subspace $\mathcal{R}(\mathcal{T}\mathcal{T}^\star) \subseteq \mathcal{R}(\mathcal{T})$. In particular, we have $\mathcal{R}(\mathcal{T}) = \mathcal{T}(\overline{\mathcal{R}(\mathcal{T}^\star)})$[1] and $\mathcal{R}(\mathcal{T}\mathcal{T}^\star) = \mathcal{T}(\mathcal{R}(\mathcal{T}^\star))$, so $\mathcal{R}(\mathcal{T}\mathcal{T}^\star)$ is generally a *strict* subset of $\mathcal{R}(\mathcal{T})$, unless $\mathcal{R}(\mathcal{T}^\star)$ is a closed set. It is well known that for ill-posed inverse problems, the operator $\mathcal{T}^\star$ generally does not have a closed range space (Carrasco et al., 2007), thus in general Assumption 2 imposes non-trivial restrictions on the ill-posedness of the inverse problem. In Section 6.4, we provide a more concrete example to illustrate these restrictions.

Source conditions are common assumptions used to derive strong convergence rate guarantees in the inverse problem literature. They have been widely used for both inverse problems with known operators (*e.g.*, Engl et al., 1996; Ito and Jin, 2014) and IV problems with unknown operators (*e.g.*, Florens et al., 2011; Carrasco et al., 2007; Liao et al., 2021). A standard source condition in the literature is that the solution h_0 satisfies $h_0 \in \mathcal{R}((\mathcal{T}\mathcal{T}^\star)^{\beta/2})$ for a positive exponent $\beta > 0$. Our source condition in Assumption 2 can be shown to be equivalent to $h_0 \in \mathcal{R}((\mathcal{T}\mathcal{T}^\star)^{1/2})$ via the spectral theory of linear operators (Cavalier, 2011). Thus, our Assumption 2 is a source condition of this kind with source exponent $\beta = 1$.

Assumption 2 implies that there exists $\bar{g}_0 \in L_2(Z)$ such that

$$r_0 = \mathcal{T}\mathcal{T}^\star \bar{g}_0. \tag{6}$$

In fact, any $\bar{g}_0$ satisfying Equation (6) is closely related to the least norm solution h_0.

1. To see this note that $L_2(X) = \overline{\mathcal{R}(\mathcal{T}^\star)} \oplus \mathcal{R}(\mathcal{T}^\star)^\perp$, and $\mathcal{R}(\mathcal{T}^\star)^\perp = \mathcal{N}(\mathcal{T})$, so $\mathcal{T}(L_2(X)) = \mathcal{T}(\overline{\mathcal{R}(\mathcal{T}^\star)})$.

Lemma 2. *If Assumption 2 holds, then $\bar{g}_0$ satisfies (6) if and only if $\mathcal{T}^\star \bar{g}_0 = h_0$.*

In particular, given Lemma 2, the functions $\bar{g}_0$ that satisfy Equation (6) are given by:

$$\mathcal{N}_{h_0}(\mathcal{T}^\star) := \{g \in L_2(Z) : \mathcal{T}^\star g = h_0\}. \tag{7}$$

In the next subsection, we will show the importance of the source condition given by Assumption 2. In particular, this condition ensures that we can obtain h_0 from the saddle points of $L(h, g)$.

4.2. Saddle Points of the Minimax Optimization

Here, we characterize the saddle points of $L(h, g)$ under Assumption 2, as follows:

Lemma 3. *Suppose Assumption 2 holds and let h_0 be the least norm solution in Equation (2) and $\mathcal{N}_{h_0}(\mathcal{T}^\star)$ be the set of functions given in Equation (7). Then, the set of saddle points of $L(h, g)$ over $h \in L_2(X), g \in L_2(Z)$, i.e., the points (h', g') that satisfy*

$$L(h, g') \geq L(h', g') \geq L(h', g), \ \forall h \in L_2(X), \forall g \in L_2(Z),$$

is given by the set $\{h_0\} \times \mathcal{N}_{h_0}(\mathcal{T}^\star) = \{(h_0, \bar{g}) : \bar{g} \in \mathcal{N}_{h_0}(\mathcal{T}^\star)\}$.

It is well-known that (h', g') is a saddle point if and only if we have the "strong duality" condition $\inf_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g) = \sup_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h, g)$ and

$$h' \in \arg\min_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g), \quad g' \in \arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h, g).$$

We provide formal proof for this in Section I. Given this equivalent characterization of the saddle point, we can obtain the following corollary from Lemma 3.

Corollary 1. *If Assumption 2 holds, then we have*

$$h_0 = \arg\min_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g), \quad \mathcal{N}_{h_0}(\mathcal{T}^\star) = \arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h, g). \tag{8}$$

It is worth noting that the equality for h_0 in Equation (8) holds even without the source condition. Moreover, the strong duality $\inf_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g) = \sup_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h, g)$ also holds in the absence of this source condition. However, the source condition is important to establish the existence of $\arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h, g)$ and the second statement in (8). Equivalently, this shows that the source condition guarantees the existence of stationary Lagrangian multipliers for the problem in Equation (2), and the set of stationary Lagrangian multipliers is given by $\mathcal{N}_{h_0}(\mathcal{T}^\star)$.

So far we have demonstrated that Assumption 2 is a sufficient condition for the existence of saddle points. Interestingly, it is also a necessary condition for their existence.

Lemma 4. *Suppose Assumption 1 that $r_0 \in \mathcal{R}(\mathcal{T})$ holds. Then, there exists a saddle point of $L(h, g)$ if and only if Assumption 2 holds.*

The above lemma is proved by first showing that the saddle point exists if and only if there exists a solution to $\arg\min_{g \in L_2(Z)} \|\mathcal{T}^* g - h_0\|_2^2$. We then demonstrate that the existence of this optimization problem is equivalent to the source condition (2). Our Lemma 3 and Lemma 4 show that

the source condition is closely related to the existence of stationary Lagrangian multipliers for the constrained optimization formulation of h_0. To our knowledge, this relation is novel in the literature.

Lemma 3 characterizes the saddle points over the unrestricted $L_2(X)$ and $L_2(Z)$ spaces. However, in practical estimation, we can only use some function classes $\mathcal{H} \subset L_2(X), \mathcal{G} \subset L_2(Z)$ with limited statistical complexity. For these two restricted classes to capture some saddle points, we need them to satisfy the following realizability assumptions.

Assumption 3 (Realizability of the least norm solution)**.** *We have* $h_0 \in \mathcal{H}$.

Assumption 4 (Realizability of the stationary Lagrange multiplier)**.** *We have* $\mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G} \neq 0$.

The realizability assumptions above require that the function classes $\mathcal{H}$ and $\mathcal{G}$ are well-specified, in that they contain at least some true saddle points. In particular, Assumption 4 is equivalent to $h_0 \in \mathcal{T}^*\mathcal{G}$. In the following theorem, we further extend the saddle point characterization of h_0 in Corollary 1 to these restricted classes under these realizability conditions.

Theorem 1 (Key identification theorem)**.** *Suppose Assumptions 2 to 4 hold. Then*

$$h_0 = \arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g).$$

Theorem 1 shows that under the source condition and the realizability assumptions, the min-max optimization of our proposed objective over the function classes $\mathcal{H}, \mathcal{G}$ can recover the saddle points in the classes. At a high level, the proof of this theorem works by showing: (1) saddle points over the original class remain saddle points over the restricted classes; (2) any additional saddle points over the restricted classes are best-responses to saddle points over the original class; and (3) h_0 is a unique best response to any $\bar{g} \in \mathcal{N}_{h_0}(\mathcal{T}^\star)$ as a result of strong convexity of $L(h, g)$ in h induced by $\langle h, h \rangle_{L_2(X)}$. See Section B for details.

5. Finite Sample Guarantees

As discussed in Section 4, our proposed minimax optimization formulation can identify the target least norm solution h_0 when the population distribution is known. In this section, we further show that our finite-sample estimator $\hat{h}_{\text{mn}}$ in Equation (4) converges to h_0, and we derive its L_2 error rate.

Theorem 2 (L_2 convergence rates)**.** *Suppose Assumptions 2 to 4 hold. Then, we have*

$$\|\hat{h}_{\text{mn}} - h_0\|_2 \leq \sqrt{2 \sup_{h \in \mathcal{H}, g \in \mathcal{G}} \left| (\mathbb{E}_n - \mathbb{E})[(Y - h(X))\, g(Z) + 0.5 h(X)^2] \right|}.$$

Notably, the assumptions required in Theorem 2 are identical to those in Theorem 1. In particular, both theorems only require that the function classes $\mathcal{H}$ and $\mathcal{G}$ satisfy the realizability conditions Assumptions 3 and 4. Realizability is a fundamental assumption in statistical learning theory. However, we can easily extend our theorem when realizability does not hold as we will later see in Theorem 3. To the best of our knowledge, existing minimax IV regression estimators additionally require much stronger conditions such as $\mathcal{T}\mathcal{H} \subset \mathcal{G}$ or $\mathcal{T}^*\mathcal{G} \subset \mathcal{H}$. These conditions are often referred to as the closedness condition, and they impose additional restrictions on the operator $\mathcal{T}$. See Section 6 for a detailed discussion.

It then remains to bound the right-hand side term in Theorem 2. This is an empirical process term, which can be easily upper-bounded by invoking standard statistical learning theory for any reasonable function classes $\mathcal{H}, \mathcal{G}$ with bounded statistical complexities. In particular, we can use standard symmetrization arguments to bound the right-hand side of Theorem 2 with the Rademacher complexities of $\mathcal{H}, \mathcal{G}$. The Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$ of class $\mathcal{H}$ is defined as $\mathfrak{R}_n(\mathcal{H}) = n^{-1}\mathbb{E}[\sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i h(X_i)]$ where $\{\sigma_1, \ldots, \sigma_n\}$ are independent random variables drawn from the Rademacher distribution. The Rademacher complexity $\mathfrak{R}_n(\mathcal{G})$ of class $\mathcal{G}$ can be defined analogously.

Corollary 2. *Suppose Assumptions 2 to 4 hold. Let $\|Y\| \le C_Y$, $\|h\|_\infty \le C_\mathcal{H}$ for any $h \in \mathcal{H}$ and $\|g\|_\infty \le C_\mathcal{G}$ for $g \in \mathcal{G}$. Then, there exists a universal positive constant c such that with probability at least $1-\delta$, we have*

$$\|\hat{h}_{\mathrm{mn}} - h_0\|_2 \le c\sqrt{(C_\mathcal{H} + C_\mathcal{G})(\mathfrak{R}_n(\mathcal{G}) + \mathfrak{R}_n(\mathcal{H})) + (C_\mathcal{G} + C_\mathcal{H})C_\mathcal{H}\sqrt{\ln(1/\delta)/n}}$$

Furthermore, for given function classes $\mathcal{H}, \mathcal{G}$, we can obtain final L_2 convergence rates by plugging in off-the-shelf results of Rademacher complexities. For example, the following corollary is obtained by instantiating Theorem 2 to finite classes.

Corollary 3. *When $\mathcal{H}, \mathcal{G}$ are finite classes, with probability at least $1-\delta$, we have $\|\hat{h}_{\mathrm{mn}} - h_0\|_2 = \mathrm{Poly}(C_\mathcal{H}, C_\mathcal{G}) \left(\frac{\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}\right)^{1/4}$ where $\mathrm{Poly}(C_\mathcal{H}, C_\mathcal{G})$ is a polynomial term in $C_\mathcal{H}$ and $C_\mathcal{G}$.*

As another example, we instantiate Theorem 2 for more general nonparametric classes whose complexity are characterized by their covering numbers.

Corollary 4. *Let $M(\epsilon, \mathcal{H}, \|\cdot\|_\infty)$ and $M(\epsilon, \mathcal{G}, \|\cdot\|_\infty)$ be covering numbers of $\mathcal{H}, \mathcal{G}$ with respect to L^∞-norm. Suppose $\ln(M(\epsilon, \mathcal{H}, \|\cdot\|_\infty)) = O(\epsilon^{-\beta})$ and $\ln(M(\epsilon, \mathcal{G}, \|\cdot\|_\infty)) = O(\epsilon^{-\beta})$ for some $\beta > 0$, and the conditions of in Corollary 2 hold. Then with probability at least $1-\delta$, we have*

$$\|\hat{h}_{\mathrm{mn}} - h_0\|_2 = \begin{cases} \mathrm{Poly}(C_\mathcal{H}, C_\mathcal{G})\{n^{-1/4} + (\ln(1/\delta)/n)^{1/4}\}, & (\beta < 2) \\ \mathrm{Poly}(C_\mathcal{H}, C_\mathcal{G})\{n^{-1/4}\ln(n) + (\ln(1/\delta)/n)^{1/4}\}, & (\beta = 2) \\ \mathrm{Poly}(C_\mathcal{H}, C_\mathcal{G})\{n^{-1/(2\beta)} + (\ln(1/\delta)/n)^{1/4}\} & (\beta > 2). \end{cases}$$

If we specialize Corollary 4 to Sobolev balls $\mathcal{H}, \mathcal{G}$ with smoothness parameter α and input dimension d, we have $\beta = d/\alpha$, so the rates become $O(n^{-1/4})$ when $\alpha/d > 2$ and $O(n^{-\alpha/(2d)})$ when $\alpha/d \le 2$. It is an interesting question whether this rate is optimal. Although Chen and Reiss (2011) derives a minimax rate for NPIV regression estimation, their result requires the NIPV equation to have a unique solution and they impose stronger conditions on the function classes, so it is not directly comparable to our rate. A thorough investigation of the rate optimality is left for future work.

Finally, we also consider the case where the function classes $\mathcal{H}, \mathcal{G}$ are misspecified so they may not satisfy the realizability assumptions. This result is useful when we use sieve estimators based on sample-dependent function classes $\mathcal{H}$ and $\mathcal{G}$, that approximate certain function spaces. For example, $\mathcal{H}, \mathcal{G}$ can be linear models with polynomial basis functions or neural networks with growing dimensions, which can gradually approach Hölder or Sobolev balls (Chen, 2007).

Theorem 3 (Finite sample result under misspecification)**.** *Suppose Assumption 2 holds, and there exists $h^\dagger \in \mathcal{H}$ and $g^\dagger \in \mathcal{G}$ such that $\|h^\dagger - h_0\|_2 \le \epsilon_h$ and $\inf_{\bar{g}_0 \in \mathcal{N}_{h_0}(\mathcal{T}^\star)} \|g^\dagger - \bar{g}_0\|_2 \le \epsilon_g$. Then*

$$\|\hat{h}_{\mathrm{mn}} - h_0\|_2 \le \sqrt{\{2C_\mathcal{H} + C_\mathcal{G}\}\epsilon_h + C_\mathcal{H}\epsilon_g + 2\sup_{h\in\mathcal{H}, g\in\mathcal{G}} |(\mathbb{E}_n - \mathbb{E})[(Y - h(X))\, g(Z) + 0.5h(X)^2]|}.$$

Compared to Theorem 2, the upper bound in Theorem 3 involves additional misspecification errors ϵ_h, ϵ_g due to misspecified $\mathcal{H}, \mathcal{G}$. The empirical process term in Theorem 3 can be again bounded by Rademacher complexities.

6. Discussions

In this section, we compare our method with existing minimax NPIV estimators in Dikkala et al. (2020); Liao et al. (2020); Bennett et al. (2022) as they are most relevant. Other existing minimax estimators are similar so we only briefly review them in Section 1.1.

6.1. Comparisons to Dikkala et al. (2020)

Dikkala et al. (2020) considers the following minimax estimator:

$$\hat{h}_{\text{pro}} = \arg\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} \tilde{L}_n(h,g), \quad \tilde{L}_n(h,g) := -0.5\mathbb{E}_n[g^2(Z)] + \mathbb{E}_n[(Y - h(X))\, g(Z)].$$

Here for simplicity, we omit possible additional regularizers for h and g.

Dikkala et al. (2020) assumes the closedness condition that $\mathcal{T}(\mathcal{H} - h_0^\diamond) \subset \mathcal{G}$ where $h_0^\diamond$ can be an arbitrary solution to $\mathcal{T}h = r_0$ (note this condition is invariant to the choice of $h_0^\diamond$). Under this condition, letting $L(h,g)$ be the population analog of $\tilde{L}_n(h,g)$, it can be shown that $\max_{g\in\mathcal{G}} \mathbb{E}[\tilde{L}(h,g)] = 0.5\mathbb{E}[(\mathcal{T}(h_0^\diamond - h)[Z])^2] = 0.5\mathbb{E}[([\mathcal{T}h](Z) - r_0(Z))^2]$. In other words, the minimax objective in Dikkala et al. (2020) is used to approximate the projected MSE objective under the closedness condition. In contrast, our proposed minimax objective is motivated by the method of Lagrange multipliers, and it does not need the closedness condition.

To compare the theory in Dikkala et al. (2020) with our theory, we consider finite classes $\mathcal{H}, \mathcal{G}$ for simplicity. Then the theory in Dikkala et al. (2020) implies that if $\mathcal{N}_{r_0}(\mathcal{T}) \cap \mathcal{H} \neq 0$ and $\mathcal{T}(\mathcal{H} - h_0^\diamond) \subset \mathcal{G}$ for $h_0^\diamond \in \mathcal{N}_{r_0}(\mathcal{T})$, then we have $\mathbb{E}[\{\mathcal{T}(\hat{h}_{\text{pro}} - h_0^\diamond)\}^2(Z)] = O\left(\left(\frac{\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}\right)^{1/2}\right)$ with probability $1-\delta$.

Note that the rate $O((\ln(|\mathcal{H}||\mathcal{G}|)/n)^{1/2})$ above is faster than our rate $O((\ln(|\mathcal{H}||\mathcal{G}|)/n)^{1/4})$ in Corollary 3. However, the rate above is for the weak projected MSE, while our rate in Corollary 3 is for the stronger L_2 error, so they are not comparable. In particular, the projected MSE rate cannot translate into an L_2 rate without further restrictions. Dikkala et al. (2020) consider restrictiting the ill-posedness measure $\sup_{h\in\mathcal{H}} \frac{\mathbb{E}[\{\hat{h}-h)\}^2(X)]}{\mathbb{E}[\{\mathcal{T}(\hat{h}-h)\}^2(Z)]}$. However, this ill-posedness measure may generally be infinite, and in fact is guaranteed to be infinite when the solutions to the NPIV problem are nonunique, so using it to get L_2 convergence rates is often problematic.

Remark 1 (Enjoy the best of both worlds)**.** Here we observe that the estimator in Dikkala et al. (2020) can achieve a fast projected MSE rate while our estimator achieves a slow L_2 rate. One may wonder whether it is possible to achieve both guarantees at the same time. We explore this question in Section A and find this is possible if we put aside computational considerations. □

6.2. Comparison to Liao et al. (2020)

Liao et al. (2020) builds on Dikkala et al. (2020) and incorporates additional Tikhonov regularization into the minimax optimization:

$$\min_{h\in\mathcal{H}}\max_{g\in\mathcal{G}} -0.5\mathbb{E}_n[g^2(Z)] + \mathbb{E}_n[(Y-h(X))\,g(Z)] + \alpha\mathbb{E}_n[h^2(X)]. \tag{9}$$

Liao et al. (2020) also needs the closedness assumption in Dikkala et al. (2020) and a realizability assumption that the Tikhonov regularized solution $h_{0,\alpha}$ is contained in $\mathcal{H}$ for small α. In addition, they assume that the NPIV solution is unique and satisfies a source condition with exponent $\beta \in (0,1]$, and the regularization strength α vanishes to 0 at an appropriate rate as $n\to\infty$. Under these conditions, they can derive an L_2 convergence rate. In particular, their L_2 rate has the order $O(n^{-1/6})$ when the function classes are *e.g.* finite or VC, and $\beta = 1$.

Our proposed estimator and theory significantly differ from Liao et al. (2020). Specifically, our estimator does not involve the $\mathbb{E}_n[g^2(Z)]$ term and our regularized term $\mathbb{E}_n[h^2(X)]$ has a constant coefficient 0.5 but Equation (9) needs a vanishing α. Moreover, our theory accommodates non-unique solutions, and uses different realizability assumptions. Notably, under our source condition $\beta = 1$, our convergence rate $O(n^{-1/4})$ is faster than the rate $O(n^{-1/6})$ in Liao et al. (2020).

6.3. Comparison to Bennett et al. (2022)

Under the same source condition, Bennett et al. (2022) [2] formulate the L_2 error of h_0 as projected MSEs: $\mathbb{E}[\{h_0-h\}^2(X)] = \mathbb{E}[(\mathcal{T}^*\{\bar{g}_0-g\})^2(X)]$ where $\mathcal{T}^*g = h$ and $\mathcal{T}^*\bar{g}_0 = h_0$. First, note that for any fixed $\bar{g}_0$ such that $T^\star\bar{g}_0 = h_0$, we have

$$\mathcal{N}_{h_0}(\mathcal{T}^*) = \arg\min_{g\in\mathcal{G}} \mathbb{E}[(\mathcal{T}^*\{\bar{g}_0-g\})^2(X)] = \arg\min_{g\in\mathcal{G}} 0.5\mathbb{E}[(\mathcal{T}^*g)^2(X)] - \mathbb{E}[Yg(Z)].$$

Then, under the closedness assumption $\mathcal{T}^*\mathcal{G}\subset\mathcal{H}$, we have

$$\mathcal{N}_{h_0}(\mathcal{T}^*) = \arg\max_{g\in\mathcal{G}}\min_{h\in\mathcal{H}} 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[\{Y-h(X)\}g(Z)].$$

Then, noting that the inner minimizer h satisfies $\mathcal{T}^*g = h$ for any given g, and recalling the original goal is to find h_0 such that $\mathcal{T}^*\bar{g}_0 = h_0$, we can deduce that [3]

$$\{\bar{g}_0, h_0\} = \arg\max_{g\in\mathcal{G}}\arg\min_{h\in\mathcal{H}} 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[\{Y-h(X)\}g(Z)].$$

Finally, their proposed estimator $\hat{h}_{\text{ffli}}$ is given by replacing expectations with empirical averages.

In comparison to our proposed estimator $\hat{h}_{\text{mn}}$, the difference lies in the flip of $\arg\max$ and $\arg\min$. Since $\mathcal{G},\mathcal{H}$ could be non-convex, the two estimators are generally different. Indeed, this results in a significant difference in terms of the required assumptions. In $\hat{h}_{\text{ffli}}$, the primary assumptions are the source condition, $\bar{g}_0\in\mathcal{G}$, and $\mathcal{T}^*\mathcal{G}\subset\mathcal{H}$ (note that $h_0\in\mathcal{H}$ is implicit from the latter two conditions). Conversely, in our proposed estimator $\hat{h}_{\text{mn}}$, the primary assumptions are the source condition and $\bar{g}_0\in\mathcal{G}, h_0\in\mathcal{H}$. This condition is strictly weaker as we dispense with the requirement of closedness. This improvement is significant due to the inherent conflict between the source condition and closedness, as elucidated next.

2. Note the main focus of Bennett et al. (2022) is to estimate the Riesz representator (in their notation, $q^\dagger$) with L_2 error rates. However, their argument is easily adapted to our scenario.
3. Here, letting a loss function to be $L^\star(h,g)$, the equation $(\bar{g}_0,h_0) = \arg\min_g\arg\max_h L(h,g)$ means $\bar{g}_0 = \arg\min_g\max_h L(h,g)$ and $h_0 = \arg\max_h L(h,\bar{g}_0)$.

6.4. Tension between Source Condition and Closedness

In Sections 6.1 to 6.3, the existing estimators all require certain closedness assumption, either $\mathcal{T}(\mathcal{H} - h_0^\diamond) \subset \mathcal{G}$ for an arbitrary solution $h_0^\diamond$ to $\mathcal{T}h = r_0$, or $\mathcal{T}^*\mathcal{G} \subset \mathcal{H}$. In contrast, our proposed estimator does not need any closedness assumption. In this subsection, we show that the closedness conditions are inherently in tension with the source condition. This illustrates the benefit of getting rid of the closedness condition. For simplicity, we consider a compact linear operator $\mathcal{T}$ that admits a singular value decomposition (SVD) $\{\sigma_i, u_i, v_i\}_{i=1}^\infty$, where $\{u_i\}_{i=1}^\infty, \{v_i\}_{i=1}^\infty$ are orthonormal bases in the Hilbert spaces $L_2(Z), L_2(X)$, respectively, and $\sigma_1 \geq \sigma_2 \geq \cdots$ are the singular values. It follows that $\mathcal{T}v_i = \sigma_i u_i, \mathcal{T}^*u_i = \sigma_i v_i$, and $\mathcal{T}\mathcal{T}^\star$ has the SVD $\{\sigma_i^2, u_i, u_i\}_{i=1}^\infty$. Here we assume a compact operator merely for a simple countable SVD. Non-compact operators can be handled similarly, but involve more cumbersome notations (Cavalier, 2011).

To understand the source condition in Assumption 2, we write the function r_0 as $r_0 = \sum_{i=1}^\infty \gamma_i u_i$ with $\sum_{i=1}^\infty \gamma_i^2 < \infty$. The source condition $r_0 \in \mathcal{R}(\mathcal{T}\mathcal{T}^\star)$ means that there exists $\bar{g}_0 = \sum_{i=1}^\infty \beta_i u_i$ with $\sum_{i=1}^\infty \beta_i^2 < \infty$ such that $h_0 = \mathcal{T}^*\mathcal{T}\bar{g}_0$. It follows from the SVD of $\mathcal{T}\mathcal{T}^\star$ that $\gamma_i = \sigma_i^2\beta_i$. Therefore, the source condition requires $\sum_{i=1}^\infty \gamma_i^2/\sigma_i^4 < \infty$. This means that the function r_0 needs to be sufficiently smooth relative to the spectrum of $\mathcal{T}$. Obviously, the source condition is more readily satisfied when the decaying rate of $\{\sigma_i\}_{i=1}^\infty$ is slower, *i.e.*, when the operators $\mathcal{T}$ and $\mathcal{T}^\star$ are less smooth. In contrast, the closedness conditions are generally more easily satisfied when $\{\sigma_i\}_{i=1}^\infty$ decays faster and the operators $\mathcal{T}$ and $\mathcal{T}^\star$ are more smooth. Hence, we observe that the source condition and closedness imply opposing restrictions on the smoothness of the operators $\mathcal{T}$ and $\mathcal{T}^\star$.

7. Conclusion

In this paper, we study NPIV regression with general function approximation. We propose a penalized minimax estimator based on a novel constrained optimization formulation of the least norm IV solution. We prove that our estimator converges to this least norm solution, and derive its L_2 convergence rate under a source condition and realizability assumptions on both function classes for the minimax estimator. Notably, our estimator does not require uniqueness of the NPIV solution, and it avoids a closedness condition commonly assumed for existing minimax estimators. There are many interesting future directions of research. One direction is extending our work to more general inverse problems, including nonlinear inverse problems (Ito and Jin, 2014). Another direction is extending our work to IV quantile regression (Chernozhukov et al., 2017).

Acknowledgement

This material is based upon work supported by the National Science Foundation under Grants No. 1846210 and 1939704. Xiaojie Mao acknowledges support from National Key R&D Program of China (2022ZD0116700) and National Natural Science Foundation of China (No. 72201150 and No. 72293561).

References

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

Donald Andrews and James H Stock. Inference with weak instruments. 2005.

Donald WK Andrews. Examples of l2-complete and boundedly-complete distributions. *Journal of econometrics*, 199(2):213–220, 2017.

Isaiah Andrews, James H. Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753, 2019. doi: 10.1146/annurev-economics-080218-025643. URL https://doi.org/10.1146/annurev-economics-080218-025643.

Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

Andrii Babii and Jean-Pierre Florens. Is completeness necessary? estimation in nonidentified linear models. *arXiv preprint arXiv:1709.03473*, 2017.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Inference on strongly identified functionals of weakly identified functions. *arXiv e-prints*, pages arXiv–2208, 2022.

Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Laurent Cavalier. Inverse problems in statistics. In *Inverse problems and high-dimensional estimation*, pages 3–96. Springer, 2011.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Qihui Chen. Robust and optimal estimation for partially linear instrumental variables models with partial identification. *Journal of Econometrics*, 221(2):368–380, 2021.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

Xiaohong Chen and Markus Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(3):497–521, 2011.

Xiaohong Chen, Victor Chernozhukov, Sokbae Lee, and Whitney K Newey. Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809, 2014.

Victor Chernozhukov, Christian Hansen, and Kaspar Wüthrich. *Instrumental variable quantile regression*. Chapman and Hall/CRC, 2017.

Timothy M Christensen. Nonparametric stochastic discount factor decomposition. *Econometrica*, 85 (5):1501–1536, 2017.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*, 2020.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Ben Deaner. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Juan Carlos Escanciano, Stefan Hoderlein, Arthur Lewbel, Oliver Linton, and Sorawoot Srisuma. Nonparametric euler equation identification and estimation. *Econometric Theory*, 2020.

Jean-Pierre Florens, Jan Johannes, and Sébastien Van Bellegem. Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*, 27(3):472–496, 2011.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

Peter Hall and Joel L Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

Joel L Horowitz. Asymptotic normality of a nonparametric instrumental variables estimator. *International Economic Review*, 48(4):1329–1349, 2007.

Joel L Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2): 347–394, 2011.

Audrey Huang and Nan Jiang. Beyond the return: Off-policy function estimation under user-specified error-measuring distributions. In *Neurips*, 2022.

Kazufumi Ito and Bangti Jin. *Inverse problems: Tikhonov theory and algorithms*, volume 22. World Scientific, 2014.

Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

Myrto Kalouptsidi, Paul T Scott, and Eduardo Souza-Rodrigues. Linear iv regression estimators for structural dynamic discrete choice models. *Journal of Econometrics*, 222(1):778–804, 2021.

Masahiro Kato, Masaaki Imaizumi, Kenichiro McAlinn, Shota Yasui, and Haruo Kakehi. Learning causal models from conditional moment restrictions by importance weighting. In *International Conference on Learning Representations*, 2021.

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

Luofeng Liao, You-Lin Chen, Zhuoran Yang, Bo Dai, Mladen Kolar, and Zhaoran Wang. Provably efficient neural estimation of structural equation models: An adversarial approach. In *Advances in Neural Information Processing Systems*, volume 33, pages 8947–8958, 2020.

Luofeng Liao, Zuyue Fu, Zhuoran Yang, Yixin Wang, Mladen Kolar, and Zhaoran Wang. Instrumental variable value iteration for causal offline reinforcement learning. *arXiv preprint arXiv:2102.09907*, 2021.

Ruiqi Liu, Zuofeng Shang, and Guang Cheng. On deep instrumental variables estimate. *arXiv preprint arXiv:2004.14954*, 2020.

Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose Blanchet. Machine learning for elliptic pdes: fast rate generalization bound, neural scaling law and minimax optimality. *arXiv preprint arXiv:2110.06897*, 2021.

Wang Miao, Lan Liu, Eric Tchetgen Tchetgen, and Zhi Geng. Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*, 2015.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020.

Whitney K Newey. Nonparametric instrumental variables estimation. *American Economic Review*, 103(3):550–56, 2013.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Andres Santos. Instrumental variable methods for recovering continuous linear functionals. *Journal of Econometrics*, 161(2):129–146, 2011.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Sheng Wang, Jun Shao, and Jae Kwang Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, pages 1097–1116, 2014.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020.

Bing Yu et al. The deep ritz method: a deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Maximum moment restriction for instrumental variable regression. *arXiv preprint arXiv:2010.07684*, 2020.

Appendix A. Enjoy the Best of Both Worlds

Thus far, we have encountered two types of guarantees: slow L_2 rates and fast projected MSEs. The next step is to obtain guarantees that possess both properties. If we put aside issues of computational efficiency, then this is actually achievable. The estimator is defined as follows:

$$\hat{h}_{\mathrm{both}} = \arg\min_{h \in \mathcal{H}_n} \max_{g \in \mathcal{G}} \tilde{L}_n(h, g)$$

where

$$\mathcal{H}_n = \{h \in \mathcal{H}; \max_{g \in \mathcal{G}} L_n(h, g) - \min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L_n(h, g) \leq \mu_n\}.$$

Here, μ_n is some hyperparameter. The set $\mathcal{H}_n$ is defined so that each of its element has the L_2 convergence guarantee under the source condition.

Theorem 4 (fast projected MSEs + slow L_2 errors)**.** *Suppose $\mathcal{H}, \mathcal{G}$ are finite for simplicity. Suppose $h_0 \in \mathcal{H}, \mathcal{T}(\mathcal{H} - h_0) \subset \mathcal{G}, \mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G} \neq \emptyset$. Then, when we take $\mu_n = (C_{\mathcal{H}} + C_{\mathcal{G}})^2 \sqrt{\ln(|\mathcal{H}||\mathcal{G}|/\delta)/n}$, with probability $1 - \delta$, we have*

$$\|\mathcal{T}(\hat{h}_{\mathrm{both}} - h_0)\|_2 \leq c(C_{\mathcal{H}} + C_{\mathcal{G}}) \sqrt{\frac{\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}}, \quad \|\hat{h}_{\mathrm{both}} - h_0\|_2 \leq c(C_{\mathcal{H}} + C_{\mathcal{G}}) \left(\frac{\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}\right)^{1/4}.$$

Appendix B. General Characterization of Saddle Points

First, notice

$$\{h_0\} = \min_{h \in L_2(X)} L(h, \bar{g}_0) \tag{10}$$

for any $\bar{g}_0 \in \mathcal{N}_{h_0}(\mathcal{T}^*)$. In other words, the optimal response to $L(h, \bar{g}_0)$ is uniquely h_0. It follows from two observations: (1) h_0 is a best response for any element $\bar{g}_0$ in $\mathcal{N}_{h_0}(\mathcal{T}^*)$, since $(h_0, \bar{g}_0)$ is a saddle point by Lemma 3; and (2) the best response for each $\bar{g}_0$ is unique, since $L(h, \bar{g}_0)$ is strictly convex in h, due to the $\langle h, h \rangle_{L_2(X)}$ term.

Next, we invoke the following general characterization of saddle points. Here, $(\tilde{x}, \tilde{y}) \in \arg\min_{x \in \mathcal{X}'} \arg\max_{y' \in \mathcal{Y}'} f(x, y)$ means $\tilde{x} \in \arg\min_{x \in \mathcal{X}'} \max_{y \in \mathcal{Y}'} f(x, y)$ and $\tilde{y} \in \arg\max_{y \in \mathcal{Y}'} f(\tilde{x}, y)$.

Lemma 5 (Characterization of saddle points over constrained sets)**.** *Let $\mathcal{Z}$ be a set of saddle points for $f(x, y)$ over $\mathcal{X}, \mathcal{Y}$. Let $\mathcal{Z}_{\mathcal{X}} = \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y), (\cdot, \tilde{\mathcal{Z}}_{\mathcal{X}}) = \arg\max_{y \in \mathcal{Y}} \arg\min_{x \in \mathcal{X}} f(x, y)$. Then, for $\mathcal{X}' \subset \mathcal{X}, \mathcal{Y}' \subset \mathcal{Y}$, if $\mathcal{Z} \cap (\mathcal{X}', \mathcal{Y}')$ is non-empty, we have*

$$\mathcal{Z}_{\mathcal{X}} \cap \mathcal{X}' \subset \arg\min_{x \in \mathcal{X}'} \max_{y \in \mathcal{Y}'} f(x, y) \tag{11}$$

and

$$\arg\min_{x \in \mathcal{X}'} \max_{y \in \mathcal{Y}'} f(x, y) \subset \tilde{\mathcal{Z}}_{\mathcal{X}} \cap \mathcal{X}'. \tag{12}$$

In Lemma 5, the primary assumption $\mathcal{Z} \cap (\mathcal{X}', \mathcal{Y}') \neq \emptyset$ means that some saddle point (with respect to $\mathcal{X}, \mathcal{Y}$) is included in $\mathcal{X}', \mathcal{Y}'$. The equation (11) states that any saddle points $(\mathcal{Z}_{\mathcal{X}} \cap \mathcal{X}')$ over unconstrained function classes $(\mathcal{X}, \mathcal{Y})$ are still saddle points over constrained function classes $(\mathcal{X}', \mathcal{Y}')$. The equation (12) states that any saddle point over constrained function classes $(\mathcal{X}', \mathcal{Y}')$ is included in $\tilde{\mathcal{Z}}_{\mathcal{X}}$.

We combine the above characterization of saddle points with Lemma 3 by setting $(\mathcal{X}', \mathcal{Y}') = (\mathcal{H}, \mathcal{G}), (\mathcal{X}, \mathcal{Y}) = (L_2(X), L_2(Z)), f(x, y) = L(h, g)$. As an immediate consequence, when $h_0 \in \mathcal{H}, \mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G} \neq \emptyset$ (i.e., saddle points are included in $(\mathcal{H}, \mathcal{G})$), using (11), we have $\{h_0\} \subset \arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g)$. Next, using (12), we have $\arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g) \subset \{h_0\}$ since $(\cdot, h_0) = \arg\max_{g \in \mathcal{G}} \arg\min_{h \in \mathcal{H}} L(h, g)$ by (10).

Appendix C. Computational Perspective

To solve the optimization problem in Equation (4), we can leverage the recent advances in minimax optimization algorithms, even when the function classes $\mathcal{H}$ and $\mathcal{G}$ are neither convex nor concave, such as neural network classes (Daskalakis et al., 2017). In particular, using a Reproducing kernel Hilbert space (RKHS) ball as $\mathcal{G}$ is particularly convenient, since then the inner maximization problem in Equation (4) has a closed form solution. Specifically, when $\mathcal{G} = \{g : \|g\|_K \leq 1\}$ for a positive definite kernel $K : D_Z \times D_Z \to \mathbb{R}$ and its associated RKHS norm $\|\cdot\|_K$, Equation (4) reduces to

$$\arg\min_{h \in \mathcal{H}} 0.5\mathbb{E}_n[h^2(X)] + \left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (Y_i - h(X_i)) K(Z_i, Z_j) (Y_j - h(X_j)) \right)^{1/2}.$$

Appendix D. Proof in Section 2

D.1. Proof of Lemma 1

Here, we have $\mathcal{N}_{r_0}(\mathcal{T}) = h_0 + \mathcal{N}(\mathcal{T})$. The least norm solution h_0 among $\mathcal{N}_{r_0}(\mathcal{T})$ is the projection of any element in $\mathcal{N}_{r_0}(\mathcal{T})$ onto (the closed subspace) $\mathcal{N}(\mathcal{T})^\perp$. Hence, $\{h_0\} = \mathcal{N}(\mathcal{T})^\perp \cap \mathcal{N}_{r_0}(\mathcal{T}) = \overline{\mathcal{R}(\mathcal{T}^\star)} \cap \mathcal{N}_{r_0}(\mathcal{T})$. Here, we use $\mathcal{N}(\mathcal{T})^\perp = \overline{\mathcal{R}(\mathcal{T}^\star)}$.

Appendix E. Proof in Section 4

E.1. Proof of Lemma 2

It is clear from Lemma 1.

E.2. Proof of Lemma 3

The proof is as follows. From Section I, a point $(h', g') \in (L_2(X), L_2(Z))$ is a saddle point if and only if the strong duality holds and $h' \in \arg\min_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g)$ and $g' \in \arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h, g)$. We check this condition.

Hence, we first show

$$\{h_0\} = \arg\min_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g), \quad 0.5\|h_0\|_2^2 = \min_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h, g) \tag{13}$$

First, for any $h \neq \mathcal{N}_{r_0}(\mathcal{T})$, we have $\sup_{g \in L_2(Z)} L(h,g) = \infty$. Hence, the solution needs to belong to $\mathcal{N}_{r_0}(\mathcal{T})$. Since $\sup_{g \in L_2(Z)} L(h,g) = 0.5\mathbb{E}[h^2(X)]$ for any $h \in \mathcal{N}_{r_0}(\mathcal{T})$, using Lemma 2, thus, from the definition of h_0, the solution is h_0.

Next, we show

$$\mathcal{N}_{h_0}(\mathcal{T}^\star) = \arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h,g), \quad 0.5\|h_0\|_2^2 = \max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h,g). \tag{14}$$

We solve the inner minimization problem first. Then,

$$\begin{aligned}
\inf_{h \in L_2(X)} L(h,g) &= \inf_{h \in L_2(X)} 0.5\|h - \mathcal{T}^\star g\|_2^2 + \langle r_0, g \rangle_{L_2(Z)} - 0.5\langle \mathcal{T}^\star g, \mathcal{T}^\star g \rangle_{L_2(X)} \\
&= \langle r_0, g \rangle_{L_2(Z)} - 0.5\langle \mathcal{T}^\star g, \mathcal{T}^\star g \rangle_{L_2(X)} \\
&= \langle \mathcal{T} h_0, g \rangle_{L_2(Z)} - 0.5\langle \mathcal{T}^\star g, \mathcal{T}^\star g \rangle_{L_2(X)} && \text{(Use } r_0 = \mathcal{T} h_0\text{)} \\
&= -0.5\|\mathcal{T}^\star g - h_0\|_2^2 + 0.5\|h_0\|_2^2.
\end{aligned}$$

By using Assumption 2, since $\mathcal{N}_{h_0}(\mathcal{T}^*)$ is not empty, we have

$$\mathcal{N}_{h_0}(\mathcal{T}^*) = \arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h,g).$$

Finally, since the strong duality holds from (13) and (14), the set of saddle points is $(h_0, \mathcal{N}_{h_0}(\mathcal{T}^\star))$.

E.3. Proof of Lemma 4

Recall the saddle point exists if and only if $\arg\min_{h \in L_2(X)} \sup_{g \in L_2(Z)} L(h,g)$ and $\arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h,g)$ exist and the strong duality holds. We already show that Assumption 2 is sufficient to ensure the existence of the saddle point. In this proof, we show Assumption 2 is necessary to ensure the existence of the saddle point.

To ensure the existence of saddle point, we need to ensure the existence of $\arg\max_{g \in L_2(Z)} \inf_{h \in L_2(X)} L(h,g)$. This optimization problem is equivalent to

$$\arg\min_{g \in L_2(Z)} \|\mathcal{T}^\star g - h_0\|_2^2 \tag{15}$$

as we see in the proof of Lemma 3. This solution exists if and only if $h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp$. To prove this, we define a projection operator onto $\overline{\mathcal{R}(\mathcal{T}^\star)}$ as $P_{\overline{\mathcal{R}(\mathcal{T}^\star)}}$. Then, the solution of (15) exists if and only if $P_{\overline{\mathcal{R}(\mathcal{T}^\star)}} h_0 \in \mathcal{R}(\mathcal{T}^\star)$. Here, $P_{\overline{\mathcal{R}(\mathcal{T}^\star)}} h_0 \in \mathcal{R}(\mathcal{T}^\star)$ implies

$$h_0 = P_{\overline{\mathcal{R}(\mathcal{T}^\star)}} h_0 + (I - P_{\overline{\mathcal{R}(\mathcal{T}^\star)}}) h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp.$$

Besides, $h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp$ implies $P_{\overline{\mathcal{R}(\mathcal{T}^\star)}} h_0 \in \mathcal{R}(\mathcal{T}^\star)$ recalling $\mathcal{H} = \overline{\mathcal{R}(\mathcal{T}^\star)} \bigoplus \mathcal{R}(\mathcal{T}^\star)^\perp$. This finishes proving that the solution of (15) exists if and only if $h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp$.

Finally, recall $h_0 \in \overline{\mathcal{R}(\mathcal{T}^\star)}$ using Lemma 1. Thus, $h_0 \in \mathcal{R}(\mathcal{T}^\star) + \mathcal{R}(\mathcal{T}^\star)^\perp$ implies $h_0 \in \mathcal{R}(\mathcal{T}^\star)$ since if $h_0 = h_{0,2} + h_{0,3}, h_{0,2} \in \mathcal{R}(\mathcal{T}^\star), h_{0,3} \in \mathcal{R}(\mathcal{T}^\star)^\perp$, we have $h_{0,3} = h_0 - h_{0,2} \in \overline{\mathcal{R}(\mathcal{T}^\star)} \cap \mathcal{R}(\mathcal{T}^\star)^\perp = \{0\}$.

The statement is concluded by the fact $h_0 \in \mathcal{R}(\mathcal{T}^\star)$ implies $r_0 \in \mathcal{R}(\mathcal{T}\mathcal{T}^\star)$.

E.4. Proof of Lemma 5

Clearly, each element in $\mathcal{Z} \cap (\mathcal{X}, \mathcal{Y})$ is a saddle point over $\mathcal{X}', \mathcal{Y}'$ since this is a saddle point over $\mathcal{X}, \mathcal{Y}$. Therefore,

$$\mathcal{Z}_{\mathcal{X}} \cap \mathcal{X}' \subset \operatorname*{arg\,min}_{x \in \mathcal{X}'} \max_{y \in \mathcal{Y}'} f(x, y).$$

Now, we prove the second statement. Let (x_0, y_0) be an element in $\mathcal{Z} \cap (\mathcal{X}, \mathcal{Y})$ (this exists and this is a saddle point). Then, take:

$$\tilde{x} \in \operatorname*{arg\,min}_{x \in \mathcal{X}'} \sup_{y \in \mathcal{Y}'} f(x, y), \quad \tilde{y} \in \operatorname*{arg\,max}_{y \in \mathcal{Y}'} \inf_{x \in \mathcal{X}'} f(x, y).$$

Since $(\tilde{x}, \tilde{y})$ is a saddle point over $\mathcal{X}', \mathcal{Y}'$, we have

$$f(x_0, y_0) \geq f(x_0, \tilde{y}) \geq f(\tilde{x}, \tilde{y}) \geq f(\tilde{x}, y_0) \geq f(x_0, y_0).$$

Then, the above inequalities are equalities. Hence, we have

$$f(x_0, \tilde{y}) = f(\tilde{x}, y_0), \quad f(x_0, y_0) = f(x_0, \tilde{y})$$

This means that

$$\tilde{x} \in \mathcal{Z}'_{\mathcal{X}} \subset \mathcal{X}'.$$

recalling $y_0 \in \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x, y)$.

E.5. Proof of Theorem 1

We show two proofs.

First Proof. We use Lemma 5. First, using (11),

$$\{h_0\} \subset \operatorname*{arg\,min}_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g).$$

Second, we use (12). Here, recalling the proof of Lemma 3, we have

$$(\mathcal{N}_{h_0}(\mathcal{T}^*), h_0)) \in \operatorname*{arg\,max}_{g \in \mathcal{G}} \operatorname*{arg\,min}_{h \in \mathcal{H}} L(h, g).$$

Therefore, using Lemma 3, we have

$$\operatorname*{arg\,min}_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g) \subset \{h_0\}.$$

Hence,

$$\operatorname*{arg\,min}_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} L(h, g) = \{h_0\}.$$

Second Proof. We give more direct proof to show the finite sample result later.

We take some element $\bar{g}_0$ from $\mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G}$. This satisfies $\mathcal{T}^* \bar{g}_0 = h_0$. We define

$$L(h, g) \coloneqq 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[g(Z)\{Y - h(X)\}],$$
$$\hat{g}(h) \coloneqq \arg\max_{g \in \mathcal{G}} L(h, g), \quad \hat{h} \coloneqq \arg\min_{h \in \mathcal{H}} \sup_{g \in \mathcal{G}} L(h, g),$$

and $\hat{g} \coloneqq \hat{g}(\hat{h})$. Hence, for any $h \in \mathcal{H}$,

$$\begin{aligned}
&L(h, \bar{g}_0) - L(h_0, \bar{g}_0) \\
&= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[\bar{g}_0(Z)\{Y - h(X)\}] - 0.5\mathbb{E}[\{h_0\}^2(X)]0\mathbb{E}[\bar{g}_0(Z)\{Y - h_0(X)\}] \\
&= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[\bar{g}_0(Z)\{h_0(X) - h(X)\}] - 0.5\mathbb{E}[\{h_0\}^2(X)] \\
&= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[h_0(X)\{h_0(X) - h(X)\}] - 0.5\mathbb{E}[\{h_0\}^2(X)] \\
&\qquad\qquad \text{(We use } \mathbb{E}[h_0(X) \mid Z] = \bar{g}_0(Z)\text{)} \\
&= 0.5\mathbb{E}[\{h(X) - h_0(X)\}^2].
\end{aligned}$$

Therefore, for any $h \in \mathcal{H}$,

$$\mathbb{E}[\{h(X) - h_0(X)\}^2] = L(h, \bar{g}_0) - L(h_0, \bar{g}_0). \tag{16}$$

Furthermore,

$$\begin{aligned}
L(\hat{h}, \hat{g}) &\geq L(\hat{h}, \bar{g}_0) && \text{(Construction of estimators)} \\
&\geq L(h_0, \bar{g}_0) && \text{(Saddle point property)} \\
&\geq L(h_0, \hat{g}(h_0)). && \text{(Saddle point property)}
\end{aligned}$$

Since we have $L(\hat{h}, \hat{g}) \leq L(h_0, \hat{g}(h_0))$ from the definition, all of the above inequalities are equalities. Then, we have

$$L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) = 0. \tag{17}$$

In conclusion, combining (16) with (17), we have

$$\mathbb{E}[\{\hat{h}(X) - h_0(X)\}^2] \leq L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) = 0.$$

Hence, $\hat{h}(X) = h_0(X)$.

Appendix F. Proof of Section 5

F.1. Proof of Theorem 2

We take some element $\bar{g}_0$ from $\mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G}$. This satisfies $\mathcal{T}^* \bar{g}_0 = h_0$. We define

$$\begin{aligned}
L(h, g) &\coloneqq 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[g(Z)\{Y - h(X)\}], \\
L_n(h, g) &\coloneqq 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n[g(Z)\{Y - h(X)\}], \\
\hat{g}(h) &\coloneqq \arg\max_{g \in \mathcal{G}} 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n[g(Z)\{Y - h(X)\}], \\
M(\mathcal{H}, \mathcal{G}) &\coloneqq \sup_{h \in \mathcal{H}, g \in \mathcal{G}} |(\mathbb{E}_n - E)[\{Y - h(X)\}g(Z) + 0.5h(X)^2]|.
\end{aligned}$$

Using (16), recall for any $h \in \mathcal{H}$, we have

$$\mathbb{E}[\{h(X) - h_0(X)\}^2] = L(h, \bar{g}_0) - L(h_0, \bar{g}_0).$$

Here, we have

$$\begin{aligned}
L_n(\hat{h}, \hat{g}(\hat{h})) &\geq L_n(\hat{h}, \bar{g}_0) && \text{(Construction of estimators)} \\
&\geq L(\hat{h}, \bar{g}_0) - M(\mathcal{H}, \mathcal{G}) \\
&\geq L(h_0, \bar{g}_0) - M(\mathcal{H}, \mathcal{G}) && \text{(Saddle point property)} \\
&\geq L(h_0, \hat{g}(h_0)) - M(\mathcal{H}, \mathcal{G}) && \text{(Saddle point property)} \\
&\geq L_n(h_0, \hat{g}(h_0)) - 2M(\mathcal{H}, \mathcal{G}) \\
&\geq L_n(\hat{h}, \hat{g}(\hat{h}) - 2M(\mathcal{H}, \mathcal{G}). && \text{(Construction of estimators.)}
\end{aligned}$$

Therefore, we have

$$L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \leq 2M(\mathcal{H}, \mathcal{G}).$$

Finally, we have

$$\mathbb{E}[\{\hat{h}(X) - h_0(X)\}^2] \leq L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \leq 2M(\mathcal{H}, \mathcal{G}).$$

F.2. Proof of Corollary 2

We calculate the following empirical process term:

$$\sup_{h \in \mathcal{H}, g \in \mathcal{G}} |(\mathbb{E}_n - \mathbb{E})[\{Y - h(X)\}g(Z) + 0.5h(X)^2]|.$$

Then, from Wainwright (2019, Theorem 4.10), this is upper-bounded by

$$c\left\{\mathfrak{R}_n(\mathcal{A}_1) + \mathfrak{R}_n(\mathcal{A}_2) + \mathfrak{R}_n(\mathcal{A}_3) + (C_\mathcal{G} + C_\mathcal{H})C_\mathcal{H}\sqrt{\ln(1/\delta)/n}\right\}$$

where

$$\mathcal{A}_1 = \{yg(z); g \in \mathcal{G}\}, \quad \mathcal{A}_2 = \{h(x)g(z); h \in \mathcal{H}, g \in \mathcal{G}\}, \quad \mathcal{A}_3 = \{0.5h(x)^2; h \in \mathcal{H}\}.$$

First, we have

$$\mathfrak{R}_n(\mathcal{A}_1) \lesssim C_\mathcal{G}\mathfrak{R}_n(\mathcal{G}).$$

Secondly, we have

$$\mathfrak{R}_n(\mathcal{A}_2) \lesssim (C_\mathcal{H} + C_\mathcal{G})(\mathfrak{R}_n(\mathcal{G}) + \mathfrak{R}_n(\mathcal{H})).$$

Here, we use the proof of Kallus et al. (2021, Proof of Corollary 3). Thirdly, we have

$$\mathfrak{R}_n(\mathcal{A}_3) \lesssim 2C_\mathcal{H}\mathfrak{R}_n(\mathcal{H}).$$

Combining all results together, the empirical process term is upper-bounded by

$$c\left\{(C_\mathcal{H} + C_\mathcal{G})(\mathfrak{R}_n(\mathcal{G}) + \mathfrak{R}_n(\mathcal{H})) + (C_\mathcal{G} + C_\mathcal{H})C_\mathcal{H}\sqrt{\ln(1/\delta)/n}\right\}.$$

F.3. Proof of Corollary 4

We combine the Dudley integral Theorem 5 with Corollary 2.

F.4. Proof of Theorem 3

We take some element $\bar{g}_0$ from $\mathcal{N}_{h_0}(\mathcal{T}^\star) \cap \mathcal{G}$. This satisfies $\mathcal{T}^* \bar{g}_0 = h_0$.

$$
\begin{aligned}
L(h, g) &:= 0.5\mathbb{E}[h^2(X)] + \mathbb{E}[g(Z)\{Y - h(X)\}], \\
L_n(h, g) &:= 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n[g(Z)\{Y - h(X)\}], \\
\hat{g}(h) &:= \arg\max_{g \in \mathcal{G}} 0.5\mathbb{E}_n[h^2(X)] + \mathbb{E}_n[g(Z)\{Y - h(X)\}], \\
M(\mathcal{H}, \mathcal{G}) &:= \sup_{h \in \mathcal{H}, g \in \mathcal{G}} |(\mathbb{E}_n - E)[\{Y - h(X)\}g(Z) + 0.5h(X)^2]|.
\end{aligned}
$$

and $\hat{g} = \hat{g}(\hat{h})$. Recall for any $h \in \mathcal{H}$,

$$
\mathbb{E}[\{h(X) - h_0(X)\}^2] \leq L(h, \bar{g}_0) - L(h_0, \bar{g}_0).
$$

Furthermore,

$$
\begin{aligned}
&L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \\
&= \underbrace{-L(h_0, \bar{g}_0) + L(h^\dagger, \hat{g}(h^\dagger))}_{(a)} \underbrace{-L(h^\dagger, \hat{g}(h^\dagger)) + L(\hat{h}, g^\dagger)}_{(c)} \underbrace{-L(\hat{h}, g^\dagger) + L(\hat{h}, \bar{g}_0)}_{(f)}.
\end{aligned}
$$

Term (a) is upper-bounded as follows:

$$
\begin{aligned}
L(h^\dagger, \hat{g}(h^\dagger)) - L(h_0, \bar{g}_0) &\leq 0.5\mathbb{E}[\{h^\dagger\}^2(X)] + \|\hat{g}(h^\dagger)\|_2 \|h_0 - h^\dagger\|_2 - 0.5\mathbb{E}[h_0^2(X)] \\
&\leq 0.5\mathbb{E}[\{h^\dagger\}^2(X)] + \sup_g \|g\|_2 \|h_0 - h^\dagger\|_2 - 0.5\mathbb{E}[h_0^2(X)] \\
&\leq 0.5\|h^\dagger + h_0\|_2 \|h^\dagger - h_0\|_2 + \{\sup_g \|g\|_2\}\|h_0 - h^\dagger\|_2 \\
&\leq \{2C_{\mathcal{H}} + C_{\mathcal{G}}\}\|h_0 - h^\dagger\|_2.
\end{aligned}
$$

Term (c) is upper-bounded as follows:

$$
\begin{aligned}
-L(h^\dagger, \hat{g}(h^\dagger)) + L(\hat{h}, g^\dagger) &\leq -L(h^\dagger, \hat{g}(h^\dagger)) + L_n(h^\dagger, \hat{g}(h^\dagger)) - L_n(h^\dagger, \hat{g}(h^\dagger)) + L_n(\hat{h}, g^\dagger) - L_n(\hat{h}, g^\dagger) + L(\hat{h}, g^\dagger) \\
&\leq M(\mathcal{H}, \mathcal{G}) + (-L_n(h^\dagger, \hat{g}(h^\dagger)) + L_n(\hat{h}, \hat{g})) + M(\mathcal{H}, \mathcal{G}) \\
&\leq 2M(\mathcal{H}, \mathcal{G}).
\end{aligned}
$$

The term (f) is upper-bounded as follows:

$$
L(\hat{h}, \bar{g}_0) - L(\hat{h}, g^\dagger) \leq \|\hat{h}\|_2 \|g_0 - g^\dagger\|_2 \leq C_{\mathcal{H}} \|g_0 - g^\dagger\|_2
$$

In conclusion, we have

$$
\mathbb{E}[\{\hat{h}(X) - h_0(X)\}^2] \leq L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \leq \{2C_{\mathcal{H}} + C_{\mathcal{G}}\}\|h_0 - h^\dagger\|_2 + C_{\mathcal{H}}\|g_0 - g^\dagger\|_2 + 2M(\mathcal{H}, \mathcal{G}).
$$

Appendix G. Proof of Section 6

G.1. Proof of Rate in Section 6.1

Recall

$$\hat{h}_{\mathrm{pro}} = \arg\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} \tilde{L}_n(h,g), \quad \tilde{L}_n(h,g) \coloneqq -0.5\mathbb{E}_n[g^2(Z)] + \mathbb{E}_n[\{Y-h(X)\}g(Z)].$$

Let

$$\hat{g}_h \coloneqq \arg\max_{g\in\mathcal{G}} \tilde{L}_n(h,g), \quad g_h \coloneqq \mathbb{E}[Y-h(X)\mid Z=\cdot],$$
$$\Gamma(h,g) \coloneqq -0.5g^2(Z) + \{Y-h(X)\}g(Z), \quad \kappa(h,g) \coloneqq \Gamma(h,g) - \Gamma(h,g_h).$$

First Step. Our goal is to show

$$\forall h\in\mathcal{H}; |\mathbb{E}_n[\kappa(h,\hat{g}_h)]| \lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}|/\delta)}{n}. \tag{18}$$

We fix h hereafter.

Here, first, we have

$$\mathbb{E}[\kappa(h,\hat{g}_h)] = 0.5\mathbb{E}[(\hat{g}_h - g_h)^2(Z)].$$

Then,

$$\begin{aligned}\mathbb{E}[\kappa(h,\hat{g}_h)] &\leq \mathbb{E}_n[\kappa(h,\hat{g}_h)] + |(\mathbb{E}-\mathbb{E}_n)[\kappa(h,\hat{g}_h)]| \\ &\leq |(\mathbb{E}-\mathbb{E}_n)[\kappa(h,\hat{g}_h)]|.\end{aligned}$$

From the first line to the second line, we use the definition of the estimator and $g_h \in \mathcal{G}$.

Now, we use Bernstein's inequality. With probability $1-\delta$, we have

$$\forall g\in\mathcal{G}, \forall h\in\mathcal{H}; (\mathbb{E}-\mathbb{E}_n)[\kappa(h,g)] \leq \sqrt{\frac{\mathrm{var}[\kappa(h,g)]\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}} + \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}.$$

In the following, we condition on this event. Then, we have

$$\mathbb{E}[\kappa(h,\hat{g}_h)] \lesssim \sqrt{\frac{\mathrm{var}[\kappa(h,\hat{g}_h)]\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}} + \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}. \tag{19}$$

Here, we have

$$\begin{aligned}\mathrm{var}[\kappa(h,\hat{g}_h)] &= \mathbb{E}[\{\Gamma(h,g_h) - \Gamma(h,\hat{g}_h)\}^2] \\ &= \mathbb{E}[0.25\{\hat{g}_h(Z) - g_h(Z)\}^2\{\hat{g}_h(Z) + g_h(Z)\}^2] \\ &\leq C_{\mathcal{G}}^2\mathbb{E}[\{\hat{g}_h(Z) - g_h(Z)\}^2].\end{aligned}$$

Therefore, combining the above with (19), we obtain

$$\mathbb{E}[\{\hat{g}_h(Z) - g_h(Z)\}^2] \lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}.$$

Hence,

$$\begin{aligned}|\mathbb{E}_n[\kappa(h,\hat{g}_h)]| &\leq |\mathbb{E}[\kappa(h,\hat{g}_h)]| + |(\mathbb{E}_n - \mathbb{E})[\kappa(h,\hat{g}_h)]| \\ &= 0.5\mathbb{E}[\{\hat{g}_h(Z) - g_h(Z)\}^2] + |(\mathbb{E}_n - \mathbb{E})[\kappa(h,\hat{g}_h)]| \\ &\lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n} + |(\mathbb{E}_n - \mathbb{E})[\kappa(h,\hat{g}_h)]| \\ &\lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}. \qquad \text{(Use (19))}\end{aligned}$$

Second Step. We define

$$\Xi(h) := \Gamma(h, g_h) - \Gamma(h_0, g_{h_0}).$$

Note $\Gamma(h_0, g_{h_0}) = 0$ since h_0. Furthermore,

$$\mathbb{E}[\Xi(h)] = \mathbb{E}[g_h^2(Z)], \quad \mathbb{E}[\Xi^2(h)] \leq (C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\mathbb{E}[g_h^2(Z)]. \tag{20}$$

Then,

$$\mathbb{E}[\Xi(\hat{h})] \leq \mathbb{E}_n[\Xi(\hat{h})] + |(\mathbb{E} - \mathbb{E}_n)[\Xi(\hat{h})]|$$

Here, using the first conclusion (18), we get

$$\begin{aligned}\mathbb{E}_n[\Xi(\hat{h})] &= \mathbb{E}_n[\Gamma(\hat{h}, g_{\hat{h}}) - \Gamma(h_0, 0)] \\ &\leq \mathbb{E}_n[\Gamma(\hat{h}, \hat{g}_{\hat{h}}) - \Gamma(h_0, \hat{g}_{h_0})] + c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n} \\ &\leq c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n}.\end{aligned}$$

From the first line to the second line, we use $h_0 \in \mathcal{H}$ and (18). From the second line to the third line, we use the construction of the estimator.

Therefore,

$$\mathbb{E}[\Xi(\hat{h})] \leq c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{G}||\mathcal{H}|/\delta)}{n} + |(\mathbb{E} - \mathbb{E}_n)[\Xi(\hat{h})]|.$$

Here, we use Bernstein's inequality. With probability $1 - \delta$, we have

$$\forall h \in \mathcal{H}; |(\mathbb{E} - \mathbb{E}_n)[\Xi(h)]| \leq \sqrt{\frac{\mathrm{var}[\Xi(h)]\ln(|\mathcal{H}|/\delta)}{n}} + c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{H}|/\delta)}{n}.$$

Hereafter, we condition on this event. Thus, using (20), we have

$$\mathbb{E}[g_{\hat{h}}^2(Z)] \leq \sqrt{\frac{g_{\hat{h}}^2(Z)\ln(|\mathcal{H}|/\delta)}{n}} + c\frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}.$$

Therefore, by some algebra, we obtain

$$\mathbb{E}[g_{\hat{h}}^2(Z)] \lesssim \frac{(C_{\mathcal{H}}^2 + C_{\mathcal{G}}^2)\ln(|\mathcal{H}||\mathcal{G}|/\delta)}{n}.$$

Appendix H. Proof of Section A

H.1. Proof of Theorem 4

We use the notation in Theorem 2. Take μ_n such that $2\mathcal{M}(\mathcal{H}, \mathcal{G}) \leq \mu_n$ holds with probabiltiy $1-\delta$. We condition on this event.

The guarantee in terms of projected MSEs is straightforward as long as h_0 is included in the confidence ball $\mathcal{H}_n$ with probability $1-\delta$ by following the proof in Theorem **??**. In fact, we have

$$
\begin{aligned}
&L_n(h_0, \hat{g}(h_0)) - \min_h L_n(h, \hat{g}(h)) = L_n(h_0, \hat{g}(h_0)) - L_n(\hat{h}, \hat{g}(\hat{h})) \\
&\leq L_n(h_0, \hat{g}(h_0)) - L_n(\hat{h}, g(\hat{h})) \\
&= L_n(h_0, \hat{g}(h_0)) - L(h_0, \hat{g}(h_0))) + L(h_0, \hat{g}(h_0))) - L(\hat{h}, g(\hat{h})) + L(\hat{h}, g(\hat{h})) - L_n(\hat{h}, g(\hat{h})) \\
&\leq L_n(h_0, \hat{g}(h_0)) - L(h_0, \hat{g}(h_0))) + L(h_0, g(h_0))) - L(\hat{h}, g(\hat{h})) + L(\hat{h}, g(\hat{h})) - L_n(\hat{h}, g(\hat{h})) \\
&\leq 2M(\mathcal{H}, \mathcal{G}) \leq \mu_n.
\end{aligned}
$$

Hence, $h_0 \in \mathcal{H}_n$.

Next, we prove the L_2 convergence guarantee. Here, for any $\hat{h}$ in the confidence ball $\mathcal{H}_n$, we have

$$
\begin{aligned}
L_n(\hat{h}, \hat{g}(\hat{h})) &\geq L_n(\hat{h}, \bar{g}_0) && \text{(Construction of estimators)} \\
&\geq L(\hat{h}, \bar{g}_0) - M(\mathcal{H}, \mathcal{G}) \\
&\geq L(h_0, \bar{g}_0) - M(\mathcal{H}, \mathcal{G}) && \text{(Saddle point property)} \\
&\geq L(h_0, \hat{g}(h_0)) - M(\mathcal{H}, \mathcal{G}) && \text{(Saddle point property)} \\
&\geq L_n(h_0, \hat{g}(h_0)) - 2M(\mathcal{H}, \mathcal{G}) \\
&\geq \min_h L_n(h, \hat{g}(h)) - 2M(\mathcal{H}, \mathcal{G}) \\
&\geq L_n(\hat{h}, \hat{g}(\hat{h})) - 2M(\mathcal{H}, \mathcal{G}) - \mu_n.
\end{aligned}
$$

Therefore,

$$
L_n(h_0, \hat{g}(h_0)) - \min_h L_n(h, \hat{g}(h)) \leq 2M(\mathcal{H}, \mathcal{G}) + \mu_n.
$$

Hence, the L_2 rate guarantee is ensured since

$$
\mathbb{E}[\{\hat{h}(X) - h_0(X)\}^2] \leq L(\hat{h}, \bar{g}_0) - L(h_0, \bar{g}_0) \leq 2M(\mathcal{H}, \mathcal{G}) + \mu_n.
$$

Appendix I. Auxiliary Lemmas

Lemma 6. *(x^*, y^*) is a saddle point of $f(x, y)$ over $(\mathcal{X}, \mathcal{Y})$ if and only if the strong duality holds and*

$$
x^* \in \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y), \quad y^* \in \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} f(x, y).
$$

Proof. Suppose (x^*, y^*) is a saddle point of $f(x, y)$ over $\mathcal{X}, \mathcal{Y}$. Then,

$$
\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} f(x, y) \leq \sup_{y \in \mathcal{Y}} f(x^*, y) \leq f(x^*, y^*) \leq \inf_{x \in \mathcal{X}} f(x, y^*) \leq \sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} f(x, y).
$$

Hence, the strong duality holds. The above inequalities are actually equalities. Therefore,

$$\inf_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}} f(x,y) = \sup_{y\in\mathcal{Y}} f(x^*,y) = f(x^*,y^*) = \inf_{x\in\mathcal{X}} f(x,y^*) = \sup_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}} f(x,y).$$

Hence, we have

$$x^* \in \arg\min_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}} f(x,y), \quad y^* \in \arg\max_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}} f(x,y).$$

Next, suppose the strong duality holds, and

$$x^* \in \arg\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} f(x,y), \quad y^* \in \arg\max_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}} f(x,y).$$

Then, we have

$$\max_{y\in\mathcal{Y}}\inf_{x\in\mathcal{X}} f(x,y) = \inf_{x\in\mathcal{X}} f(x,y^*) \le f(x^*,y^*) \le \sup_{y\in\mathcal{Y}} f(x^*,y) = \min_{x\in\mathcal{X}}\sup_{y\in\mathcal{Y}} f(x,y).$$

Finally, using the strong duality, the above is actually equality. Hence,

$$\inf_{x\in\mathcal{X}} f(x,y^*) = f(x^*,y^*), \quad \sup_{y\in\mathcal{Y}} f(x^*,y) = f(x^*,y^*).$$

This implies (x^*, y^*) is a saddle point since

$$\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}; f(x,y) \ge f(x^*,y^*) \ge f(x^*,y).$$

□

Theorem 5 (Dudley integral)**.** *Consider a function class $\mathcal{F}$ containing $f : \mathcal{X} \to \mathbb{R}$. Then, we have*

$$\mathcal{R}_n(\mathcal{F}) \le \inf_{\epsilon\ge 0}\left\{4\epsilon + 12\int_{\epsilon}^{\|\mathcal{F}\|_\infty}\sqrt{\frac{\ln\mathcal{N}(\tau,\mathcal{F},\|\cdot\|_\infty)}{n}}\mathrm{d}\tau\right\}.$$