Semiparametric Estimation of Signaling Games with Equilibrium Refinement

Kyoo il Kim*
Michigan State University

September 23, 2018

Abstract

This paper studies an econometric modeling of a signaling game where one informed player may have multiple types. For this game the problem of multiple equilibria arises and we achieve the uniqueness of equilibrium using an equilibrium refinement, which enables us to identify the model parameters. We then develop an estimation strategy that identifies the payoffs structure and the distribution of types from the observed actions. In this game, the type distribution is nonparametrically specified and we estimate the model using a sieve conditional MLE. We achieve the consistency and the asymptotic normality for the structural parameter estimates.

Keywords: Signaling Game, Equilibrium Refinement, Sieve Estimation
JEL Classification: C13, C14, C35, C62, C73

1 Introduction

The econometric modeling of strategic interactions has been of significant interest. Most of the existing work in the literature of static games has focused on the simultaneous move games. In these games, it is typically assumed that there is only one type of players or there is no information friction on the types of players. Therefore, the game players are symmetric in terms of information content they have on their opponents.

In some situations, however, a player may have more information than others and can take an advantage of this information asymmetry or friction. The used car market is one natural example of this sort where sellers have more information on true qualities of cars being sold. The labor market is another example of information asymmetry where a job candidate knows her true productivity type while employers do not. A vast of economic models have been developed to study these information asymmetry problems (e.g. the seminal work by Spence (1974)). Although this

*Corresponding author: Kyoo il Kim, kyookim@msu.edu; Department of Economics, Michigan State University; Tel. +1 517-353-9008.

1 Some of these models assume the game players have complete information (Bjorn and Vuong (1984, 1985), Bresnahan and Reiss (1990, 1991), Tamer (2003), Bajari, Hong and Ryan (2010)) and others assume incomplete information (Brock and Durlauf (2001, 2003), Seim (2006), Sweeting (2009), Aradillas-Lopez (2010, 2012), Bajari, Hong, Krainer, Nekipelov (2010), Wan and Xu (2014)).

2 Riley (2001) provides an extensive survey on these topics.
information asymmetry problem has been well noted both theoretically or empirically, a formal structural econometric analysis of this asymmetry problem appears only in a few studies (e.g. Ackerberg (2003), Brown (2002), Bergman (2018)).

This paper provides a structural econometric modeling of a signaling game for which an informed player moves first by taking a costly action (signaling) in the sense that she cannot be sure about the outcome of the game. The informed player can achieve the outcome she desires only when the second player (uninformed player) makes a correct assessment on the informed player’s type. Therefore, modeling beliefs of players on the distribution of types plays an important role in this game, which makes it distinct from other existing econometric game models.

One of difficulties for econometric implementation of a signaling game has been the problem of multiple equilibria or even absence of any equilibrium if we restrict strategies of players to pure strategies, which results in the nonexistence of a well-defined likelihood function over the entire set of observable outcomes. This paper shows that the non-existence problem can be resolved by allowing for a sort of mixed strategy (semi-separating in Bayesian extensive form game) and the multiplicity problem can be resolved by adopting an equilibrium refinement such as intuitive criterion by Cho and Kreps (1987), which enables us to achieve the uniqueness of equilibrium.

Even when these problems are overcome, it is generally hard to set up a flexible econometric model that is capable of dealing with a generic family of signaling games with flexible information structure. This paper takes a first step in econometric modeling of a signaling game with two players where one player, informed player, has one of two types. In particular, we develop an estimation strategy that identifies the payoffs structure and the distribution of types from the observed actions of players.

To consider a more realistic setting, we introduce non-strategic public signals about the types to the model, which can not be manipulated by the informed player, or at least the player has no incentive to do so. These signals are observed by all players and econometricians. Under the separating equilibrium, the action of the informed player reveals the true type and thus the uninformed party has no incentive to use this additional information. However, under the pooling equilibrium, the uninformed party will update her belief on types using these public signals following a Bayes’ rule. Therefore the probability of being a particular type may depend on the player’s observable characteristics.

In our approach we specify the distribution of non-strategic signals on types nonparametrically and thus estimate the primitives of the game model, such as payoffs structure and distribution of types, using a sieve conditional MLE (a conditional MLE version of Wong and Severini (1991) or Shen (1997)) where the infinite dimensional parameters are approximated by sieves. We then provide some asymptotic properties of the proposed estimator.

The organization of the paper is as follows. Section 2 describes the game model. In Section 3, we
characterize the equilibria of the game and show how we can achieve the uniqueness of equilibrium using a refinement method. In Section 4, we discuss identification and estimate the model using a sieve conditional MLE. The consistency and the asymptotic normality of the structural parameter estimates are provided. We then run a small-scale Monte Carlo experiment in Section 5. We conclude in Section 6. Technical details and mathematical proofs are gathered in the appendix.

2 Description of the Game Model

The signaling game we consider here models a situation that each player observes actions of all players (Player 1 and Player 2) including her own but one of the players has uncertainty about payoffs that are affected by types of the other player. Without loss of generality, we will assume Player 1 is informed of her type but Player 2 does not know Player 1’s type. What makes a signaling game intriguing is that Player 2 (uninformed party or receiver) controls the action and Player 1 (informed party or sender) controls the information. The receiver has an incentive to try to deduce the sender’s type from the sender’s signal, and the sender may have an incentive to mislead the receiver.

The solution concept we use for this game is a perfect Bayesian equilibrium (PBE), instead of a sequential equilibrium (SE). The former is simpler than the latter. We do so without loss of generality since the game of interest here is a finite Bayesian extensive game with observable actions, for which the sequential equilibrium is equivalent to a perfect Bayesian equilibrium in the sense that they induce the same behavior and beliefs. The formal definition of a PBE can be found in Osborne and Rubinstein (p.233, 1994).

Figure 1 illustrates the structure of the game we study, named Game G. We design this game such that it is simple but still contains all essences of a signaling game. This is an econometric modeling of the beer-quiche game or bar fighting game in Cho and Kreps (1987). In this game, we have two players. Player 1 has either of two types \{strong, weak\} with the probability of being the strong type equal to \(p\) and knows her type. After observing her type, Player 1 moves first sending one of two messages \{B, Q\} to Player 2 where \(B\) stands for “beer” and \(Q\) stands for “quiche”. Here it is known to public that beer is associated with the strong type and quiche is associated with the weak type. Also there exists a mimicking cost of pretending to be a false type. After observing Player 1’s signal, Player 2 chooses one action out of \{F, NF\} where \(F\) stands for “fight” and \(NF\) for “not fight”. Here Player 2 is more likely to choose to fight (not to fight) if the player thinks Player 1 is a weak type (strong type). After the play, payoffs are realized according to actions chosen by the two players as assigned at the terminal nodes of the game tree in Figure 1. The payoffs of this game have the following properties that are generally shared in signaling games.

- The payoffs of Player 2 (uninformed party) given her action are determined by the type of Player 1 (informed party), not by Player 1’s action (signal).
- Given Player 2’s action, each type of Player 1 has bigger payoffs by choosing the signal corresponding to her true type. “\(B\)” is the signal for the strong type and “\(Q\)” is the signal

---

4 Note that the reverse statement is not true in general (see page 234-235, Osborne and Rubinstein (1994)). However, Fudenberg and Tirole (1991) note that for a finite Bayesian extensive game with two types or with two periods, every PBE is also equivalent to a sequential equilibrium.
for the \textit{weak} type by construction (i.e., the mimicking costs are positive).

- The \textit{strong} type of Player 1 may have an incentive to signal its true type.
- The \textit{weak} type of Player 1 may have an incentive to mislead Player 2.

Many relevant strategic interactions can fit into this framework. Think of a job-market signaling in the labor market or a strategic advertising spending in the differentiated products market as leading examples. In a job-market signaling case Player 1 is a job applicant of either high or low productivity. Player 2 is a potential employer who wants to hire an applicant with high productivity. Note that high productivity types may have an incentive to undertake a costly education as signaling e.g., obtaining an advanced degree. Here the high types have this incentive because it is more costly for the low types to obtain an advanced degree (and this fact is also known to Player 2, the employer). In the \textit{beer-quiche} game framework, Player 1’s undertaking of higher education corresponds to choosing “\textit{Beer}” and Player 2’s hiring decision corresponds to the action of “\textit{NF}”.

In a strategic advertising case Player 1 is a firm whose product quality is either high or low. Engaging in costly advertising corresponds to the signaling of strong type, like having “\textit{Beer}” in the \textit{beer-quiche} game framework. Player 2’s are the consumers who want to single out the high quality product. Buying a product of the firm corresponds to the action of “\textit{NF}” in the \textit{beer-quiche} game. Here firms with high quality goods have an incentive of costly advertising because it is more costly for firms with low quality goods to engage in excessive advertising (and this fact is also known to Player 2, the consumers).

As summarized in Figure 1 the \textit{beer-quiche} game can be fully characterized by the seven structural objects:

![Figure 1. Structure of the Signaling Game](image-url)
• $u_{1s}$: payoffs difference between “$NF$” and “$F$” outcomes for the strong type of Player 1
• $u_{1w}$: payoffs difference between “$NF$” and “$F$” outcomes for the weak type of Player 1
• $\phi_{1s}$: mimicking cost of the strong type
• $\phi_{1w}$: mimicking cost of the weak type
• $u_{2s}$: payoffs difference choosing between “$NF$” and “$F$” for Player 2 when Player 1 is strong
• $u_{2w}$: payoffs difference choosing between “$NF$” and “$F$” for Player 2 when Player 1 is weak
• $p$: distribution of types

We further elaborate on the payoffs structure. Note that because an equilibrium of a game is characterized by payoffs differences of actions of Players, not by levels we take payoffs differences as the structural objects. In other words we normalize the payoffs of “$F$” equal to zero (before we account for the mimicking costs of Player 1). These normalizations and incentive parameters have some important consequences in the payoffs displayed in Figure 1. Because of these normalizations and because the payoffs of each player are determined by Player 2’s action and Player 1’s type, we only need to specify 4 payoffs differences (2 players & 2 types) between “$NF$” and “$F$”: $u_{1s}, u_{1w}, u_{2s},$ and $u_{2w}$ in Figure 1. This means 8 payoffs out of 16 payoffs are set to zero when there is no mimicking cost of each type of Player 1. In the game we also introduce a mimicking cost of Player 1 for each type, denoted by two parameters: $\phi_{1s}$ and $\phi_{1w}$. Then because the mimicking costs appear in each type of Player 1’s payoffs (when one type mimics the other type) regardless of Player 2’s action, these costs appear in four places (two for the strong type choosing “$Q$” and the other two for the weak type choosing “$B$”) of Player 1’s payoffs.

In our econometric analysis we want to identify these seven structural objects that characterize the game from the observed outcomes of actions. It is obvious that if they are fully nonparametrically specified, we cannot identify all of these objects, unless there are exclusion or parametric restrictions since we have only three independent conditional probabilities out of four possible outcomes $\{B&NF, B&F, Q&NF, Q&F\}$ that can be used for estimation.

To ensure the identification, we take the following parametric specifications of the payoffs as in Figure 2. Our detailed strategy of identification is discussed in Section 4.1.

- $u_{1s} = X_1' \beta_1 - \varepsilon_1$, $u_{1w} = X_1' \beta_1 - \varepsilon_1$
- $u_{2s} = X_2' \beta_2 - \varepsilon_2 + \phi_{2s}$, $u_{2w} = X_2' \beta_2 - \varepsilon_2 - \phi_{2w}$

Here $(X_1, X_2)$ are observable factors that affect Player 1&2’s payoffs and $(\varepsilon_1, \varepsilon_2)$ in the payoffs are unobservables to econometricians while Player 1&2 observe all of them. The constants $\phi_{2s}$ and $\phi_{2w}$ in the Player 2’s payoffs measure the degree of Player 2’s incentive to single out a particular type of Player 1 (i.e., payoffs from judging the Player 1’s type correctly). For example, $\phi_{2s}$ is the extra payoffs to Player 2 by taking the “$NF$” action (believing Player 1 is the strong type) when indeed Player 1 is the strong type.\(^5\) Throughout this paper, we assume $\phi_{2s}$ and $\phi_{2w}$ are not negative. Note that this assumption is innocuous in the sense that “strong” and “weak” types are just labels: unless we give some structures to them. By imposing $\phi_{2s} \geq 0$, we mean that Player 2 is

\(^5\)Here we assume that $\phi_{1s}, \phi_{1w}, \phi_{2s},$ and $\phi_{2w}$ are constant but we may extend the model such that these parameters also depend on characteristics of Player 1 and Player 2, respectively. This, however, makes the identification of the model rely on more restrictive conditions such as functional form or exclusion restrictions.
more likely to be better off by singling out the strong type for the action “NF” and to be better off by singling out the weak type for the action “F”.

The game tree after the parametrization in Figure 2 tells that if Player 1 is the strong type and if Player 1 chooses “Q” and Player 2 chooses “NF”, Player 1 gets $X_1^1\beta_1 - \varepsilon_1 + \phi_{1s}$ and Player 2 gets $X_2^2\beta_2 - \varepsilon_2 + \phi_{2s}$, respectively. If Player 1 is the strong type and if Player 1 chooses “Q” and Player 2 chooses “F”, they earn $-\phi_{1s}$ and 0, respectively. Other payoffs can be read similarly. Note that each player’s payoffs are not only determined by her own action but also by the other player’s action or type, which depicts the strategic interactions between the two players.

Figure 2. Parametrization of Payoffs

Let $X_1 = (1, \tilde{X}_1)$, $X_2 = (1, \tilde{X}_2)$, $\tilde{X} = (\tilde{X}_1, \tilde{X}_2) \in \mathbb{R}^{k_1} \times \mathbb{R}^{k_2}$, $X = (X_1, X_2)$, and let $\varepsilon = (\varepsilon_1, \varepsilon_2) \in \mathbb{R}^2$. We denote the vectors of parameters as $\beta_1 \in \mathbb{R}^{k_1+1}$, $\beta_2 \in \mathbb{R}^{k_2+1}$, $(\phi_{1s}, \phi_{1w}, \phi_{2s}, \phi_{2w}) \in (0, \infty)^2 \times [0, \infty)^2$, and $p \in (0, 1)$ and collect the parameters as $\theta = (\beta_1, \beta_2, \phi_{1s}, \phi_{1w}, \phi_{2s}, \phi_{2w})$. Note that this stochastic payoffs game is different from the deterministic payoffs game in that by adding the unobserved heterogeneity $(\varepsilon_1, \varepsilon_2)$, we allow that players with the same observed characteristics $(X_1, X_2)$ can produce different outcomes of the game even when they adopt the same equilibrium strategy. We also note that the symmetric payoffs structure of the game $G$ is somewhat restrictive. For example, under this structure, the following two payoffs (to be precise, the payoffs differences between the outcomes of “NF” and “F”) are the same. One is the payoff of the strong type Player 1 when she chooses “B” and Player 2 plays “NF”. The other is the payoff of the weak type Player 1 when she chooses “Q” and Player 2 chooses “NF”. This structure may be justified in some situations but can be restrictive in general. In the supplementary appendix, we relax this restriction and show that the game with asymmetric payoffs is very similar to the symmetric case in terms of equilibrium characterization.
Before we characterize the equilibria of the game, first we formalize the information structure of Players and make some stochastic assumptions.

### 2.1 Information Structure

Note that in the signaling game the true type of Player 1 is only known to Player 1 herself and Player 1 signals her type to Player 2, so Player 2 has uncertainty on her payoffs, which depends on Player 1’s type.

**Assumption 2.1 (IS)**

1. Player 1 knows her true type but Player 2 knows only the distribution of Player 1’s types (i.e., $p$ is known to Player 2).
2. Realizations of $(X_1, \varepsilon_1)$ and $(X_2, \varepsilon_2)$ are perfectly observed by both Player 1 and Player 2.
3. $(\varepsilon_1, \varepsilon_2)$ are pure shocks commonly observed by Player 1 and Player 2. They are jointly independent of any other variables in the game. They are also independent of the type of Player 1.
4. Players’ actions and beliefs constitute a Perfect Bayesian Equilibrium. Whenever there exist multiple equilibria, only one equilibrium is chosen out of these according to some equilibrium refinement.\(^6\) Players are assumed to play actions and hold beliefs about this unique equilibrium.

### 2.2 Stochastic Assumptions

We impose the following distributional assumptions on the random variables of the game. We first consider the simplest structure and generalize it later.

**Assumption 2.2 (SA)**

1. $\varepsilon_1$ and $\varepsilon_2$ are continuously distributed, statistically independent of $X$.
2. The CDF’s of $\varepsilon_1$ and $\varepsilon_2$ are continuous and denoted by $G_1(\varepsilon_1)$ and $G_2(\varepsilon_2)$ with corresponding density functions $g_1(\varepsilon_1)$ and $g_2(\varepsilon_2)$, respectively. The density functions do not depend on the structural parameter $(\theta, p)$ nor on the type of Player 1.
3. Both $\tilde{X}_1$ and $\tilde{X}_2$ can be continuous or discrete random variables. Both $\tilde{X}_1$ and $\tilde{X}_2$ are independent of the type of Player 1. We denote the density of $\tilde{X}_1$ and $\tilde{X}_2$ as $f_{\tilde{X}_1}(\cdot)$ and $f_{\tilde{X}_2}(\cdot)$, respectively. Neither $f_{\tilde{X}_1}(\cdot)$ or $f_{\tilde{X}_2}(\cdot)$ depends on the structural parameter $(\theta, p)$.

By imposing Assumption SA.1 we ensure that Player 2’s equilibrium beliefs are constructed being conditioned on the variables observed by the econometrician. We will strengthen these stochastic assumptions to ensure the validity of the econometric modeling in later sections.

\(^6\)Note that the generic uniqueness of the PBE is ensured by some refinement of the equilibrium concept.
3 Refinement and Uniqueness of Equilibrium

In this section we characterize the equilibria of the game under PBE and then show we can achieve the uniqueness of equilibrium using a refinement of Cho and Kreps (1987).\footnote{Possible disadvantage of this approach noted in the literature is that there is neither generally accepted or empirically testable procedure to determine which equilibrium will be played among multiple equilibria, especially for the simultaneous move games. We, however, note that an equilibrium selection of a signaling game is more acceptable, at least theoretically in the literature, which may justify our approach.} We introduce additional notation. Let $u_i(t_1; A_1, A_2)$ denote the payoffs of player $i \in \{1, 2\}$ when the true type of Player 1 is $t_1 \in \{t_s \equiv \text{strong}, t_w \equiv \text{weak}\}$, Player 1 chooses action (signal) $A_1 \in \{B, Q\}$ and Player 2 chooses action $A_2 \in \{NF, F\}$. We also let $A_{1t_1}$ denote the action taken by a particular type $t_1$ of Player 1. Note $p = \Pr(t_1 = t_s)$ is the prior belief of Player 2 on the type of Player 1. This is also the population distribution of types. We let $E_1[u_1(t_1; A_1, A_2)]$ be the expected payoff of Player 1 with type $t_1$ based on Player 1’s information. $\mu_2(t_1|A_1)$ denotes the posterior belief of Player 2 on the type of Player 1 after Player 2 observes the action (signal) of Player 1, $A_1$. We also let $Y_{2|A_1}$ be an indicator function that has the value 1 when Player 2 chooses the action “NF” after observing the signal $A_1$. Similarly, $A_{2|A_1}$ denotes the action of Player 2 after observing $A_1$. Finally we let $(A_1, A_2 : q)$ denote the observed actions of Player 1&2 with probability $q$ in a separating equilibrium and we let $(A_1, A_2 : 1) = (A_1, A_2)$ in a pooling equilibrium case. In the supplementary appendix, we characterize the PBE for all realizations of $(\varepsilon_1, \varepsilon_2)$ given $X$ and determine the regions of $(\varepsilon_1, \varepsilon_2) \in \mathbb{R}^2$, under which each PBE (Pooling with Beer or Quiche, Separating, Semi-separating when only one of two types plays mixing strategy) exists. Figure 3 summarizes the results and some detailed discussions follow.

Figure 3 describes the distinctive regions of $(\varepsilon_1, \varepsilon_2)$ where a particular equilibrium is supported in terms of observed actions, assuming $\phi_{1s} > 0, \phi_{1w} > 0, \phi_{2s} \geq 0, \phi_{2w} \geq 0$ (but not $\phi_{2s} = \phi_{2w} = 0$).\footnote{To define a PBE, we also need to specify out-of-equilibrium belief that supports a particular equilibrium. A specific out-of-equilibrium belief for each PBE of the game can be found in the supplementary appendix. Here we do not present such out-of-equilibrium belief to make discussion simple.} Note that Figure 3 is drawn for the case $\phi_{1w} > \phi_{1s}$ as an illustration but this is not necessary. Each rectangular figure of different colors depicts the regions of $(\varepsilon_1, \varepsilon_2)$ under which a particular observed action profile is supported by the corresponding PBE. This characterization based on the observed action profile is useful to generate the model choice probabilities for estimation later. The overlaps of the rectangular figures indicate the existence of multiple equilibria when a pooling equilibrium exists with another equilibrium at the same region. We further discuss this below.

In Figure 3 (also Figure 4 below) we need to distinguish an equilibrium from its realized outcomes. For example, in the region $S_1 \equiv \{(\varepsilon_1, \varepsilon_2)|\varepsilon_1 \in \mathbb{R} \text{ and } \varepsilon_2 > X_2^0\beta_2 + \phi_{2s}\}$, we have one equilibrium where Player 1 plays the separating equilibrium with $(A_{1s}, A_{1w}) = (B, Q)$ and Player 2 chooses $F$. However, in terms of realized outcomes, we have two possible outcomes in this region. With probability $p$ (when Player 1 is the strong type), we will observe the $(B, F)$ outcome but we can also observe the $(Q, F)$ outcome with probability $1-p$ (when Player 1 is the weak type). Likewise even though we have only one equilibrium (semi-separating where the strong type plays $B$ and the weak type mixes between $B$ and $Q$) in the region $S_2 \equiv \{(\varepsilon_1, \varepsilon_2)|\varepsilon_1 < X_1^0\beta_1 - \phi_{1w} \text{ and } X_2^0\beta_2 + p(\phi_{2s} + \phi_{2w}) - \phi_{2w} < \varepsilon_2 < X_2^0\beta_2 + \phi_{2s}\}$, we can observe all four possible outcomes because of the mixing strategy.
As it is expected from the structure of the game, under $\varepsilon_2 > X'_2\beta_2 + \phi_{2w}$ (when relatively large values $\varepsilon_2$ are realized), Player 2 is better off by choosing $F$ regardless of Player 1’s type or action and will choose $F$. Therefore, Player 1 will choose a signal corresponding to her type since, given Player 2’s action, Player 1 is better off by choosing the signal that corresponds to her true type. Similarly, under $\varepsilon_2 < X'_2\beta_2 - \phi_{2w}$ (when relatively small values of $\varepsilon_2$ are realized), Player 2 is better off by choosing $NF$ no matter what and thus under this region, Player 1 is willing to reveal her true type.

We can find similar patterns for the Player 1 when the realized values of $\varepsilon_2$ are in the moderate range. Due to the game structure, when $\varepsilon_1$ is relatively large, Player 1 tends to prefer $F$ while she prefers $NF$ when $\varepsilon_1$ is small. However, Player 1 cannot choose Player 2’s action of $F$ or $NF$ but she can only induce Player 2 to choose $F$ or $NF$ by sending proper signals. As a consequence, in Figure 3 and Figure 4 below (see the regions $S_4$ and $S_5$), we observe that when $\varepsilon_1$ is relatively large, Player 1 tends to choose $Q$ to induce $Fighting$ while she tends to choose $B$ when $\varepsilon_1$ is relatively small, to induce non-$Fighting$. This well demonstrates the key feature of a signaling game such that Player 2 (uninformed party or receiver) controls the action and Player 1 (informed party or sender) controls the information.

In Figure 3 and 4, under the region $S_3 \equiv \{(\varepsilon_1,\varepsilon_2)|X'_1\beta_1 - \phi_{1w} < \varepsilon_1 < X'_1\beta_1 + \phi_{1w}$ and $X'_2\beta_2 - \phi_{2w} < \varepsilon_2 < X'_2\beta_2 + \phi_{2w}\}$, we have the separating equilibrium (i.e., the strong type chooses $B$ and the weak type choose $Q$). This region gets larger as $\phi_{1w}$, $\phi_{1w}$, $\phi_{2w}$, and $\phi_{2w}$ become larger. This implies that when the cost of mimicking and the Player 2’s incentive to single out a particular type of Player 1 are higher, the region that supports this separating equilibrium becomes larger.
We note that in many empirical studies, researchers tend to focus on the separating equilibrium where Player 1 reveals her true type and Player 2 chooses different actions according to different types of Player 1 by imposing relevant conditions or by simply asserting a separating equilibrium is more realistic. However, Figure 3 illustrates that other kinds of equilibria can arise depending on the realizations of \((\varepsilon_1, \varepsilon_2)\). For example, in the region \(S_2 \equiv \{(\varepsilon_1, \varepsilon_2) | \varepsilon_1 < X'_1\beta_1 - \phi_{1w} \) and \(X'_2\beta_2 + p(\phi_{2s} + \phi_{2w}) - \phi_{2w} < \varepsilon_2 < X'_2\beta_2 + \phi_{2s}\}\), we have a semi-separating equilibrium where the strong type plays \(B\) and the weak type mixes between \(B\) and \(Q\). Similarly, in the region \(S_6 \equiv \{(\varepsilon_1, \varepsilon_2) | \varepsilon_1 > X'_1\beta_1 + \phi_{1s} \) and \(X'_2\beta_2 - \phi_{2w} < \varepsilon_2 < X'_2\beta_2 + p(\phi_{2s} + \phi_{2w}) - \phi_{2w}\}\), we have another semi-separating equilibrium where the weak type plays \(Q\) and the strong type mixes between \(B\) and \(Q\).

The following two theorems state the existence of PBE for all realizations of \((\varepsilon_1, \varepsilon_2)\) given \(X\) and provide conditions under which the uniqueness of equilibrium is achieved by an equilibrium refinement (Theorem 3.2).

**Theorem 3.1 (Existence of Equilibrium)**

Suppose Assumptions IS and SA hold. Suppose that \(\phi_{1s} > 0, \phi_{1w} > 0, \phi_{2s} \geq 0, \) and \(\phi_{2w} \geq 0\) (but not \(\phi_{2s} = \phi_{2w} = 0\)). Then, there exist PBE for all realizations of \((\varepsilon_1, \varepsilon_2)\) given \(X\).

See the supplementary appendix for the proof in which we determine the regions of \((\varepsilon_1, \varepsilon_2) \in \mathbb{R}^2\) under which each possible PBE exists given \(X\). These regions cover the whole \(\mathbb{R}^2\), so PBE exists for all possible realizations of \((\varepsilon_1, \varepsilon_2)\).

We note that there are regions where multiple equilibria arise in Figure 3. For these cases, using the refinement of Cho and Kreps (1987), we can achieve the uniqueness of equilibrium as we show below. In Figure 3 there exist three regions where we have multiple equilibria. In the region \(M_1 \equiv \{(\varepsilon_1, \varepsilon_2) | \varepsilon_1 > X'_1\beta_1 + \phi_{1w} \) and \(X'_2\beta_2 + p(\phi_{2s} + \phi_{2w}) - \phi_{2w} < \varepsilon_2 < X'_2\beta_2 + \phi_{2s}\}\), we have two possible equilibria. One is a pooling equilibrium \((A_{1t}, A_{1w}) = (B, B)\) with \(A_{2|B} = F\) and the other is a pooling equilibrium \((A_{1t}, A_{1w}) = (Q, Q)\) with \(A_{2|Q} = F\). In the region \(M_2 \equiv \{(\varepsilon_1, \varepsilon_2) | \varepsilon_1 < X'_1\beta_1 - \phi_{1s} \) and \(X'_2\beta_2 - \phi_{2w} < \varepsilon_2 < X'_2\beta_2 + p(\phi_{2s} + \phi_{2w}) - \phi_{2w}\}\), we also have two possible equilibria. One is a pooling equilibrium with \(Q\) and \(A_{2|Q} = NF\) and the other is a pooling equilibrium with \(B\) and \(A_{2|B} = NF\).

In Appendix A, using the intuitive criterion of Cho and Kreps (1987), we show that the pooling with \(B\) does not survive the refinement in \(M_1\) and the pooling with \(Q\) does not survive the refinement in \(M_2\). Figure 4 illustrates the uniqueness of equilibrium based on this refinement result. The intuitive criterion is based on the concept of “equilibrium dominance” which tells that a certain type should not be expected to play a certain strategy. For example, in the region \(S_4 \equiv \{(\varepsilon_1, \varepsilon_2) | \varepsilon_1 > X'_1\beta_1 + \phi_{1s} \) and \(X'_2\beta_2 + p(\phi_{2s} + \phi_{2w}) - \phi_{2w} < \varepsilon_2 < X'_2\beta_2 + \phi_{2s}\}\), Player 1 can play a pooling equilibrium strategy with \(B\) or \(Q\). However, for the pooling equilibrium with \(B\) under \(S_4\), it is not reasonable to believe that a deviation is played by the strong type because she is better off by choosing \(B\) no matter what Player 2 chooses (see Appendix A). Thus, Player 2 will conclude a deviation is played by the weak type for sure when she observes \(Q\). Under this refined belief, now we can eliminate the pooling with \(B\) in the region \(S_4\).
We summarize this uniqueness result.

**Theorem 3.2 (Uniqueness of Equilibrium)**
Suppose Assumptions IS and SA hold and that \( \phi_{1a} > 0, \phi_{1w} > 0, \phi_{2a} \geq 0, \) and \( \phi_{2w} \geq 0 \) (but not \( \phi_{2a} = \phi_{2w} = 0 \)). Suppose each Player plays an equilibrium that survives the refinement of Cho and Kreps (1987) when there exist multiple equilibria. Then, there exists unique equilibrium for each realization of \((\varepsilon_1, \varepsilon_2)\) given \(X\).

This uniqueness result helps us to recover the game parameters from the observed actions because in this case we can generate the model choice probabilities of players’ actions given a distributional assumption on the unobservables \((\varepsilon_1, \varepsilon_2)\), which should be one-to-one with the observed choice probabilities. Note that one of key challenges in the econometric implementation of empirical games has been the presence of multiple equilibria, because with multiplicity there are a set of model choice probabilities that can be all consistent with the observed ones. Below we briefly discuss an estimation strategy based on the uniqueness result and fully develop our estimation method in Section 4.

---

of Player 1 choose \( B \) and Player 2 does not fight if she observes \( B \) while she fights if she observes \( Q \) with out-of-equilibrium belief \( \mu_2(t_s|Q) \leq 0.5 \). In the other equilibrium, both types of Player 1 choose \( Q \) and Player 2 chooses not to fight if she observes \( Q \) while Player 2 fights if she observes \( B \) with out-of-equilibrium belief \( \mu_2(t_s|B) \leq 0.5 \). This example of Cho and Kreps (1987) corresponds to the region \( S_5 \equiv \{(\varepsilon_1, \varepsilon_2)|\varepsilon_1 < X_1^t_1 \beta_1 - \phi_{1w} \) and \( X_2^t_2 \beta_2 - \phi_{2w} < \varepsilon_2 < X_2^t_2 \beta_2 + p(\phi_{2a} + \phi_{2w}) - \phi_{2w}\} \) in Figure 3 and 4.
3.1 Estimation: an Overview

The uniqueness result enables us to obtain a well-defined likelihood function for estimation after some necessary normalizations. Note that in the payoffs of Player 1 the constant term $\beta_{1,0}$ in $\beta_1$ is not separately identified from $\phi_{1s}$ and $\phi_{1w}$ because in the payoffs of Player 1 we only identify $\beta_{1,0} + \phi_{1s}$ and $\beta_{1,0} - \phi_{1w}$ as the constant terms. We need to normalize either $\beta_{1,0} = 0$ or consider the symmetric case $\phi_{1} \equiv \phi_{1s} = \phi_{1w}$. We take the latter approach for illustration here but other normalization is also possible. By the same reason the constant term $\beta_{2,0}$ in $\beta_2$ is not separately identified from $\phi_{2s}$ and $\phi_{2w}$, so we let $\phi_2 \equiv \phi_{2s} = \phi_{2w}$ for the Player 2’s payoffs. Note, however, that these normalizations are not required to obtain the uniqueness of equilibrium result in Theorem 3.2.

Let $Y_1 = 1\{\text{Player 1 chooses } B\}$ and $Y_2 = 1\{\text{Player 2 chooses } NF\}$ where $1\{\cdot\}$ denotes an indicator function. From the result of Theorem 3.2, we can define the conditional probabilities of four possible observed outcomes $P_{jl}(X, \theta, p) = \Pr(Y_1 = j, Y_2 = l|X)$ for $j, l \in \{0, 1\}$ where we let $\theta = (\beta_1, \beta_2, \phi_1, \phi_2)$. The specific forms of these conditional probabilities are provided in Appendix B.

Based on these conditional probabilities, one can estimate the game parameter $\theta$ and $p$, using the conditional ML method with $n$ observations of the game outcomes as

$$\hat{\theta}_{CML, \hat{p}_{CML}} = \arg\max_{\theta \in \Theta, p \in (0, 1)} \frac{1}{n} \sum_{i=1}^{n} l(y_i|x_i, \theta, p)$$

where $l(y_i|x_i, \theta, p) = y_{i1}y_{i2}\log P_{11}(x_i, \cdot) + y_{i1}(1 - y_{i2})\log P_{10}(x_i, \cdot) + (1 - y_{i1})y_{i2}\log P_{01}(x_i, \cdot) + (1 - y_{i1})(1 - y_{i2})\log P_{00}(x_i, \cdot)$. The estimator $\hat{\theta}_{CML, \hat{p}_{CML}}$ will be consistent and asymptotically normal under standard regularity conditions (e.g. Newey and McFadden (1994)). Note that this conditional ML estimator is identical to the ML estimator since the density function of $X$ does not depend on the parameter $(\theta, p)$ by Assumption SA.3.

3.2 Public Information about the Type

Here we consider possible public signals about the type of Player 1. We denote such signals by $Z$, which are observable to all players of the game and to econometricians. In Assumption SA, we have assumed that $X_1$ is independent of the type of Player 1. However, it is likely that at least some of observed characteristics of Player 1 will reveal information regarding the type of Player 1. In the beer or quiche game story, characteristics of Player 1 such as muscle intensity, height, or age will tell how likely Player 1 is the strong type. Thus, $Z$ may include all or subset of variables $X_1$.

For the public signals, we require that Player 1 cannot strategically choose the signals $Z$ when a game is played (for example, in the beer-quiche game, Player 1 cannot work out her muscle when the game is played) or at least Player 1 does not have an incentive to do so. This means that in the game, only the action $A_1$ plays the role of the strategic signal. The public signals $Z$ will have a mixing distribution with a mixing probability $p$ and we denote the density of $Z$ by $f_Z(z) = pf_{(s)}(z) + (1 - p)f_{(w)}(z)$ where $f_{(s)}$ and $f_{(w)}$ denote the densities of $Z$ for the strong type and the weak type, respectively and $(f_{(s)}, f_{(w)}, p)$ are known to the players of the game. Player 2 has an incentive to use these signals while playing the game. When players play a separating equilibrium, these additional signals have no additional information for Player 1’s type since Player
1’s type is perfectly revealed from her action. When the players play a pooling equilibrium, Player 2 will use these additional signals to update her belief on Player 1’s type using the Bayes’ rule. Therefore, under a separating equilibrium, we have $\mu_2(t_1 = t_s | A_1, Z) = \mu_2(t_1 = t_s | A_1)$ while under a pooling equilibrium, we have

$$
\mu_2(t_1 = t_s | A_1, Z = z) = \frac{pf_s(z)}{pf_s(z) + (1 - p) \, f(w)(z)}
$$

which is also the conditional probability of being strong type given $Z = z$. Let $p(z)$ denote this conditional probability, $Pr(t_1 = t_s | Z = z) = \frac{pf_s(z)}{pf_s(z) + (1 - p) f(w)(z)}$. Thus, we have $p(Z) = \mu_2(t_1 = t_s | A_1, Z)$ under a pooling equilibrium. Note that the equation (1) and $p(z)$ become the prior $p$, when $f_s(z) = f(w)(z)$ (no mixture). We augment the information structure and the stochastic assumptions when the public signals are available as follows.

**Assumption 3.1 (IS-A)**

1 Assumption IS holds.

2 The public signals $Z$ about the types of Player 1 are known to both Player 1 and Player 2.

**Assumption 3.2 (SA-A)**

1 $\epsilon_1$ and $\epsilon_2$ are continuously distributed, statistically independent of $X$ and $Z$.

2 The CDF’s of $\epsilon_1$ and $\epsilon_2$ are continuous and denoted by $G_1(\epsilon_1)$ and $G_2(\epsilon_2)$ with corresponding density functions $g_1(\epsilon_1)$ and $g_2(\epsilon_2)$, respectively on their support $\mathbb{R}$. The density functions do not depend on the model parameters $(\theta, p)$ nor on the type of Player 1.

3 Both $\tilde{X}_1$ and $\tilde{X}_2$ can be continuous or discrete random variables. Both $\tilde{X}_1 - Z$ and $\tilde{X}_2$ are independent of the type of Player 1.$^{12}$ We denote the density of $\tilde{X}_1 - Z$ and $\tilde{X}_2$ as $f_{\tilde{X}_1 - Z}()$ and $f_{\tilde{X}_2}()$, respectively. Neither $f_{\tilde{X}_1 - Z}()$ or $f_{\tilde{X}_2}()$ depends on the structural parameter $(\theta, p)$.

4 $Z$ is random variables with the mixing density $f_Z(z) = pf_s(z) + (1 - p)f(w)(z)$. Neither $f_s(z)$ or $f(w)(z)$ depends on the structural parameter $(\theta, p)$.

In the game $G$ under Assumptions IS-A and SA-A, it is not difficult to see that we will obtain exactly the same equilibria as under Assumptions IS and SA except for replacing $\mu_2(t_1 = t_s | A_1)$ with $\mu_2(t_1 = t_s | A_1, Z = z)$. This means that when Player 2’s belief is used in the characterization of equilibrium, we have only to replace $p$ with $p(z)$. Applying the refinement of Cho and Kreps (1987) to the game with public signals $Z$, we also obtain the same uniqueness of equilibrium result as Theorem 3.2.

For our estimation below we will use a logistic specification for the posterior belief of Player 2 under a pooling equilibrium. Note that the equation (1) can be rewritten as

$$
p(z) = \frac{pf_s(z)}{pf_s(z) + (1 - p) \, f(w)(z)} = \frac{(p/(1 - p)) \, (f_s(z)/f(w)(z))}{1 + (p/(1 - p)) \, (f_s(z)/f(w)(z))}
$$

$^{12}$With some abuse of notation, $\tilde{X}_1 - Z$ denotes variables that are contained in $\tilde{X}_1$ but are excluded from $Z$. We allow $\tilde{X}_1 - Z$ to be empty and in this case the assumptions on $\tilde{X}_1 - Z$ become irrelevant.
and thus, the belief only depends on $p$ and the ratio $f_s(z)/f_w(z)$. The relationship in (2) means that for updating beliefs, we do not need to know $f_s$ and $f_w$ individually but only the ratio of the two is required. Therefore, by defining $h^o(z) = \log (f_s(z)/f_w(z))$, we can rewrite (1) as a logistic specification with $L(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$,

$$
p(z) = \frac{\exp(\log(p/(1-p)) + h^o(z))}{1 + \exp(\log(p/(1-p)) + h^o(z))} = L(\log(p/(1-p)) + h^o(z)) = L(h(z)) 
$$

recalling that the posterior belief of Player 2 under a pooling equilibrium equals to the conditional probability of being strong type given $Z = z$. We will let $h(z) = \log(p/(1-p)) + h^o(z)$ and this becomes one of our parameters of interest as $p(z) = L(h(z))$. Finally, note that we allow a subset of $Z$ or all $Z$ can be discrete. In this case a part of $h(Z)$ or $h(Z)$ can be parametric.

**Remark 1** Suppose the probability of being a strong type $p(z, \eta)$ also depends on other characteristics of Player 1, denoted by $\eta$, which are unobservable to Player 2. Then, the posterior belief of Player 2 will be still $p(z)$ such that $p(z) = E[p(Z, \eta)|Z = z]$ while the probability of being a strong type will be $p(z, \eta)$. This makes our problem more difficult in characterizing the choice probabilities implied by the model and may require integration over $\eta$.

### 4 Estimation

In this section, we develop our identification and estimation strategy. The information structure **IS-A** is maintained but the stochastic assumptions are strengthened to further facilitate estimation. To preserve the uniqueness of equilibrium, we also maintain the assumption that each player plays only one equilibrium that survives the refinement of Cho and Kreps (1987) when there exist multiple equilibria.

We introduce additional notation. Let $d_l$ denote the dimension of a vector $l$. Let $Y_1$ denote the indicator function that has the value one when Player 1 chooses “$B$” such that $Y_1 \equiv \mathbf{1}\{A_1 = B\}$. Similarly, we let $Y_2 \equiv \mathbf{1}\{A_2 = NF\}$ and let $Y = (Y_1, Y_2)$. Let $C(\cdot), C_1(\cdot), C_2(\cdot)$, and so on denote generic positive constants or functions. For a positive number $k$, we let $\lfloor k \rfloor$ denote the largest integer smaller than $k$. We also let the upper case stand for a random variable and the lower case stand for a realization of it. We use the subscript “$0$” to denote the true value of parameters. Throughout the paper, we assume that econometricians observe realizations of $X_1 \cup Z$, $X_2$, $Y_1$, and $Y_2$ but do not observe $\varepsilon_1$ or $\varepsilon_2$. We let $\gamma = (\gamma_1, \gamma_2)'$ denote parameters that determine the distributions of $\varepsilon_1$ and $\varepsilon_2$ such that $G_1(\varepsilon_1; \gamma_1)$ and $G_2(\varepsilon_2; \gamma_2)$. As in the usual discrete choice models, some normalizations of $\gamma_1$ and $\gamma_2$ are necessary. Whenever it is necessary, we will adopt such normalizations. To make notation simple, we consider the symmetric payoffs and also assume that all $\gamma$ are normalized or known in what follows. Finally, we let $W \equiv X \cup Z \in W$ and let $\alpha = (\theta, h) \in \mathcal{A} \equiv \Theta \times \mathcal{H}$.

For estimation we strengthen Assumption **SA-A** by adding a few “smoothness” conditions on $G_1(\cdot)$ and $G_2(\cdot)$. We let $\varepsilon_1$, $\varepsilon_2$, $X_1$, $X_2$, and $Z$ denote the supports of $\varepsilon_1$, $\varepsilon_2$, $X_1$, $X_2$, and $Z$, respectively. Conditions below are standard regularity conditions in the literature.

**Assumption 4.1 (SA2)**

1. $\varepsilon_1$ and $\varepsilon_2$ are continuously distributed, statistically independent of each other and of $W$. 

14
2 The CDF’s of $\varepsilon_1$ and $\varepsilon_2$ are continuous and denoted by $G_1(\varepsilon_1)$ and $G_2(\varepsilon_2)$ with corresponding density functions $g_1(\varepsilon_1)$ and $g_2(\varepsilon_2)$, respectively on their support $\mathbb{R}$. The density functions do not depend on the model parameter $\alpha$ or on the type of Player 1.

3 $G_1(\varepsilon_1)$ and $G_2(\varepsilon_2)$ are three times continuously differentiable with bounded derivatives everywhere in $\mathcal{E}_1 = \mathcal{E}_2 = \mathbb{R}$.

4 Both $\bar{X}_1$ and $\bar{X}_2$ can be continuous or discrete random variables. Both $\bar{X}_1 - Z$ and $\bar{X}_2$ are independent of the type of Player 1. Neither $f_{\bar{X}_1-Z}(\cdot)$ or $f_{\bar{X}_2}(\cdot)$ depends on the model parameter $\alpha$.

5 The support of $Z$ is compact. The density function $f_Z(z) = pf_{(s)}(z) + (1-p)f_{(w)}(z)$ is bounded and bounded away from zero. Neither $f_{(s)}(z)$ or $f_{(w)}(z)$ depends on the structural parameter $(\theta, p)$.

4.1 Identification

In this section we discuss identification of the model parameters from the observed outcomes. Let $\mathcal{L}_{Y_1Y_2}(W, \alpha)$ denote the conditional probabilities of the observed outcomes $\mathcal{L}_{y_1y_2}(W, \alpha) \equiv \Pr(Y_1 = y_1, Y_2 = y_2|W, \alpha)$ for $y_1, y_2 \in \{0, 1\}$, which are defined in Appendix B. Let $\alpha_0 \equiv (\theta_0, h_0)$ denote the true value of $\alpha$. We show that the following condition is sufficient for identification of $\alpha_0$.

**Assumption 4.2 (ID)** If $\alpha \neq \alpha_0$ for $\alpha, \alpha_0 \in \mathcal{A}$, then $\Pr(\mathcal{L}_{y_1y_2}(W, \alpha) \neq \mathcal{L}_{y_1y_2}(W, \alpha_0)) > 0$ for $y_1, y_2 \in \{0, 1\}$.

The condition says for identification to hold the model choice probabilities, evaluated at different values of parameters $\alpha$, must be different for non-negligible values of $W$. Define the population criterion function

$$Q(\alpha) \equiv Q(\theta, h) = E[l(Y|W, \theta, h)]$$

where $l(Y|W, \theta, h)$ denotes the single observation conditional log likelihood function as $l(y|w, \cdot) = y_1y_2\log \mathcal{L}_{11}(w, \cdot) + y_1(1 - y_2)\log \mathcal{L}_{10}(w, \cdot) + (1 - y_1)y_2\log \mathcal{L}_{01}(w, \cdot) + (1 - y_1)(1 - y_2)\log \mathcal{L}_{00}(w, \cdot)$. The following lemma shows that if Assumption ID holds, then $l(Y|W, \alpha)$ satisfies an information inequality result which is useful to prove the consistency of the proposed estimator $\tilde{\alpha}_n$ in (15) below.

**Lemma 4.1 (Identification)** Suppose Assumptions IS-A and SA2 hold. Further suppose Assumption ID holds. Then, $Q(\alpha) < Q(\alpha_0)$ for all $\alpha \neq \alpha_0 \in \mathcal{A}$.

Lemma 4.1 states that $\alpha_0$ is the unique maximizer of the population criterion function. Therefore Assumption ID delivers a clear idea about the identification condition. It is, however, a high level condition. Below we consider more primitive sufficient conditions that can satisfy Assumption ID using an identification at infinity argument as commonly used in the literature (e.g., Tamer (2003), Berry and Tamer (2006), and Bajari, Hong, and Ryan (2010)).

Now suppose $X_1$ and $X_2$ satisfy a full rank condition, respectively. Let $\bar{X}_{1,1}$ denote the first element of $\bar{X}_1$ and $\beta_{1,1}$ be the coefficient on $\bar{X}_{1,1}$. Without loss of generality, further assume $\beta_{1,1} > 0$ where the distribution of $\bar{X}_{1,1}$ conditional on $(\bar{X}_{1,-1} \cup Z, \bar{X}_2)$ has an everywhere positive Lebesgue density where we define $\bar{X}_{1,-1} = (\bar{X}_{1,2}, \ldots, \bar{X}_{1,k_1})'$. Further assume $\bar{X}_{1,1}$ is not included in $Z$.

Consider the model choice probabilities when $\bar{X}_{1,1}$ tends to infinity.
As $\tilde{X}_{1,1} \to \infty$, $g_1(\cdot) \to 0$ and we obtain
\[
\lim_{\tilde{X}_{1,1} \to \infty} \Pr(Y_1 = 1, Y_2 = 0|W) = p(Z) \left(1 - G_2 \left(X_2' \beta_2 + \phi_2\right)\right)
\]
\[
\lim_{\tilde{X}_{1,1} \to \infty} \Pr(Y_1 = 0, Y_2 = 1|W) = (1 - p(Z)) G_2 \left(X_2' \beta_2 - \phi_2\right).
\]

In the above the conditional choice probabilities (in the LHS) are known from the data. Then the identification is achieved if there is an unique solution in $(p(Z), \beta_2, \text{ and } \phi_2)$ to the above equations. Therefore, we can identify $p(Z)$ (and thus $h(Z) = L^{-1}(p(Z))$), $\beta_2 = (\beta_{2,1}, \ldots, \beta_{2,k_2})$, and $\beta_{2,0} + \phi_2$ from (5) since $Z \cap X_2 = \emptyset$ (inherent exclusion between Player 1 and Player 2) and similarly we can identify $p(Z), \tilde{\beta}_2$, and $\beta_{2,0} - \phi_2$ from (6). Then from $\beta_{2,0} + \phi_2$ and $\beta_{2,0} - \phi_2$ we also separately identify $\beta_{2,0}$ and $\phi_2$. To see how $p(Z)$ is identified, note that $p(Z)$ is obtained by fixing $X_2$ and varying $Z$ in (5) or (6). Once we identify $p(Z)$, now we can identify other parameters by fixing $Z$ and varying $X_2$ in (5) and (6). For example, given $p(Z)$ the identification of $\tilde{\beta}_2$ and $\beta_{2,0} + \phi_2$ is achieved because
\[
\Pr\left[\lim_{\tilde{X}_{1,1} \to \infty} \Pr\left(Y_1 = 1, Y_2 = 0|W, \tilde{\beta}_2, \beta_{2,0} + \phi_2\right) \neq \lim_{\tilde{X}_{1,1} \to \infty} \Pr\left(Y_1 = 1, Y_2 = 0|W, \tilde{\beta}_2, (\beta_{2,0} + \phi_2)^*\right)\right] > 0
\]
for any $\tilde{\beta}_2^* \neq \tilde{\beta}_2$ or $(\beta_{2,0} + \phi_2)^* \neq \beta_{2,0} + \phi_2$ since $X_2$ satisfies a full rank condition.

Once we identify $\beta_2, \phi_2,$ and $p(\cdot)$, we can treat them as known when identifying other parameters. Next, we show that the parameters of Player 1’s payoffs are also identified by combining outcomes such that the problem becomes a single agent decision problem.

**[II] Combine $(B, NF)$ and $(B, F)$ outcomes and obtain**

\[
\Pr(Y_1 = 1|W) = A(Z, X_2) G_1(X_1' \beta_1 + \phi_1) + B(Z, X_2) G_1(X_1' \beta_1 - \phi_1) + C(Z, X_2)
\]

where $A(Z, X_2), B(Z, X_2),$ and $C(Z, X_2)$ are some known functions of $p(Z)$ and $G_2(\cdot)$’s. Now we see that the identification problem of $\beta_1$ and $\phi_1$ becomes a nonlinear estimation problem of (8), whose identification conditions are standard in the literature (see e.g. Newey and McFadden (1994)). Note that $\tilde{\beta}_1 = (\beta_{1,1}, \ldots, \beta_{1,k_1}), \beta_{1,0} + \phi_1,$ and $\beta_{1,0} - \phi_1$ are identified if the vector that stacks
\[
\frac{\partial \Pr(Y_1 = 1|W)}{\partial \beta_1} = \{A(Z, X_2) g_1(X_1' \beta_1 + \phi_1) + B(Z, X_2) g_1(X_1' \beta_1 - \phi_1)\} \widetilde{X}_1
\]
\[
\frac{\partial \Pr(Y_1 = 1|W)}{\partial (\beta_{1,0} + \phi_1)} = A(Z, X_2) g_1(X_1' \beta_1 + \phi_1)
\]
\[
\frac{\partial \Pr(Y_1 = 1|W)}{\partial (\beta_{1,0} - \phi_1)} = B(Z, X_2) g_1(X_1' \beta_1 - \phi_1)
\]
satisfies a full rank condition such that there does not exist a linear functional relationship among the elements of the vector. Then from \( \beta_{1,0} + \phi_1 \) and \( \beta_{1,0} - \phi_1 \) we also separately identify \( \beta_{1,0} \) and \( \phi_1 \). Combining [I1] and [I2], we conclude that all the parameters \( (\beta_1, \beta_2, \phi_1, \phi_2, h(\cdot)) \) are identified. We summarize the result.

**Proposition 4.1** Suppose the model choice probabilities are given as in Appendix B. Suppose (i) \( X_1 \) contains at least one variable (say \( \tilde{X}_{1,1} \)) that is excluded from both \( Z \) and \( X_2 \), whose distribution has the full support on \( \mathbb{R} \); (ii) \( X_1 \) and \( X_2 \) satisfy a full rank condition, respectively; (iii) \( \beta_{1,1} > 0 \) where the distribution of \( \tilde{X}_{1,1} \) conditional on (\( \tilde{X}_{1,-1} \cup Z, \tilde{X}_2 \)) has an everywhere positive Lebesgue density; (iv) The vector that stacks (9)-(11) satisfies a full rank condition. Then, \( \beta_1, \beta_2, \phi_1, \phi_2, \) and \( h(\cdot) \) (i.e., \( p(\cdot) \)) are identified.

**Remark 2** Note that \( \beta_1, \beta_2, \phi_1, \) and \( \phi_2 \) are identified up to some normalizations of the distribution parameter \( \gamma \) as in the usual discrete choice models. For example, if \( G_1(\cdot) \) and \( G_2(\cdot) \) are the CDF’s normal distributions, we normalize them to be the CDF of standard normal distribution.

**Remark 3** We may let \( \phi_1 \) and \( \phi_2 \) depend on the characteristics of Player 1 and Player 2, respectively (e.g., some or all of \( X_1 \) and \( X_2 \)). If we let \( \phi_1 = \exp(X'_1\lambda_1) \) and \( \phi_2 = \exp(X'_2\lambda_2) \), a similar investigation in line with [I1]-[I2] reveals that we can also identify \( \lambda_1 \) and \( \lambda_2 \) due to the functional form restrictions. If we take \( \phi_1 = X'_1\lambda_1 \) and \( \phi_2 = X'_2\lambda_2 \), a similar investigation in line with [I1]-[I2] also reveals that we cannot identify \( \lambda_1 \) and \( \lambda_2 \) without suitable normalizations.

**Remark 4** We can also consider asymmetric payoffs such that \( \tilde{\beta}_1 \) is equal to \( \tilde{\beta}_s \) for the strong type and equal to \( \tilde{\beta}_w \) for the weak type. A similar investigation in line with [I1]-[I2] reveals that \( \tilde{\beta}_s \) and \( \tilde{\beta}_w \) can be also identified.

One may think that the identification of \( \beta_1 \) and \( \phi_1 \) in [I2] relies on the nonlinearity of \( G_1(\cdot) \) and \( G_2(\cdot) \). Here we illustrate that this is not the case. For this exercise only, assume that \( G_1(\cdot) \) and \( G_2(\cdot) \) are the CDF’s of uniform distributions over \([0,1]\). In this case one can show that the equation (8) becomes

\[
\Pr(Y_1=1|W) = A(Z) + B(Z)X_1\beta_1 + C(Z)\phi_1
\]

where \( A(Z), B(Z), \) and \( C(Z) \) are known functions of \( p(Z) \) and \( \phi_2 \). Then, the identification of \( \beta_1 \) and \( \phi_1 \) becomes a regression problem. Therefore, unless \( A(Z), B(Z)X_1, \) and \( C(Z) \) have a linear functional relationship, the identification result holds in this case too.

### 4.2 Sieve Estimation

We estimate the parameters of interest using a conditional sieve MLE approach. We approximate the unknown function \( h \) using sieves. For this purpose, we need to restrict the space of functions \( \mathcal{H} \) where the true function \( h_0 \) belong to. We consider the Hölder space \( \Lambda^{\nu_1}(Z) \) with order \( \nu_1 > 0 \) (see e.g., Ai and Chen (2003) and Chen (2007)).\(^{13}\) The Hölder ball (with a radius \( C_1 \)) \( \Lambda^{\nu_1}_{C_1}(Z) \) is defined accordingly as \( \Lambda^{\nu_1}_{C_1}(Z) \equiv \{ g \in \Lambda^{\nu_1}(Z) : ||g||_{\Lambda^{\nu_1}} \leq C_1 < \infty \} \) where \( || \cdot ||_{\Lambda^{\nu_1}} \) denotes the Hölder norm.

\(^{13}\)The Hölder space is a space of functions \( g : Z \rightarrow \mathbb{R} \) such that the first \( \nu_1 \) derivatives are bounded and the \( \nu_2 \)-th derivatives are Hölder continuous with the exponent \( \nu_1 - \nu_2 \in (0, 1] \), where \( \nu_2 \) is the largest integer smaller than \( \nu_1 \).
In the literature, it is well known that functions in $\Lambda^{p_1}_1(Z)$ can be well approximated by various sieves such as power series, Fourier series, splines, and wavelets. We let $\mathcal{H} = \Lambda^{p_1}_1(Z)$. In particular, we approximate $h(\cdot)$ by power series. According to Theorem 8, p.90 in Lorentz (1986) (Also see Timan (1963)), if a function $f$ is $s$-times continuously differentiable, then there exists a $K$-vector $\gamma_K$ and a triangular array of polynomials $R^K(z)$ on the compact set $Z$ such that

$$\sup_{z \in Z} |f(z) - R^K(z)^\gamma_K| \leq C_1 K^{-s/d_z}. \quad (12)$$

As as approximation of $\mathcal{H}$, we then consider a sieve space, $\mathcal{H}_n$, based on the tensor-product power series:

$$\mathcal{H}_n = \{ h(z) | h(z) = R^{K_n}(z)^\pi \text{ for all } \pi \text{ satisfying } \|h\|_{A^{p_1}} \leq C_1 \} \quad (13)$$

such that $\mathcal{H}_n \subseteq \mathcal{H}_{n+1} \subseteq \ldots \subseteq \mathcal{H}$.

Then, from (12), we can find an approximate function $h_n(\cdot) = R^{K_n}(\cdot)^\pi_{K_n} \in \mathcal{H}_n$ such that

$$\sup_{z \in Z} |h_0(z) - R^{K_n}(z)^\pi_{K_n}| = O \left( K_n^{-1/d_z} \right) \cdot \text{This leads us to obtain the estimator } \hat{\theta}_n \text{ and } \hat{\tilde{h}}_n \text{ by maximizing the following sample criterion function} \hat{\alpha}_n \equiv \left( \hat{\theta}_n, \hat{\tilde{h}}_n \right) = \max_{(\theta, h) \in \mathcal{A}_n = \Theta \times \mathcal{H}_n} \hat{Q}(\theta, h) \equiv \frac{1}{n} \sum_{i=1}^{n} \left( y_i | f_{i} \right), \text{ with } \epsilon_n = o(1). \quad (14)$$

We call $\alpha_n$ the exact sieve conditional ML estimator. However, because the complexity (in the sense defined in Appendix C.3) of the sieve space $\mathcal{A}_n$ increases with $n$ and because the maximizer of (14) is often obtained numerically, we do not require the maximization of $\hat{Q}(\theta, h)$ over $\mathcal{A}_n$ to be exact. An approximated estimator $\hat{\alpha}_n$ is enough for the asymptotic results we desire, given by

$$\hat{Q}(\hat{\alpha}_n) \geq \sup_{\alpha \in \mathcal{A}_n} \hat{Q}(\alpha) - O_p(\epsilon_n), \text{ with } \epsilon_n = o(1). \quad (15)$$

We call $\hat{\alpha}_n$ in (15) the approximate sieve ML estimator. We choose the order of $\epsilon_n$ such that it can justify desirable asymptotic results.

### 4.2.1 Consistency and Convergence Rates of the Sieve Conditional ML

In this section we show the consistency of our sieve conditional ML estimator and derive its convergence rates. The consistency of sieve MLE was derived in Wong and Severini (1991) and Geman and Hwang (1982) for i.i.d data. Some consistency results of sieve M-estimators can be found in Gallant (1987) and Gallant and Nychka (1987). Chen (2007) provides a consistency result of sieve extremum estimators allowing for non-compact infinite-dimensional $\mathcal{A}$, which is an extension of Theorem 2.1 in Newey and McFadden (1994) and of Lemma A1 in Newey and Powell (2003). Below using Theorem 3.1 in Chen (2007), we establish the consistency under a pseudo metric $\| \cdot \|_s$ defined by $\| \alpha_1 - \alpha_2 \|_s = \| \theta_1 - \theta_2 \|_E + \| h_1 - h_2 \|_\infty \text{ where } \| \cdot \|_E \text{ is the Euclidean norm and } \| h \|_\infty = \sup_{z \in Z} |h(z)|$.

The Hölder space becomes a Banach space when endowed with the Hölder norm:

$$\|g\|_{A^{p_1}} = \sup_{z} |g(z)| + \max_{a_1 + a_2 + \ldots + a_d \neq 0} \sup_{z, z' \in Z} \frac{|\nabla^a g(z) - \nabla^a g(z')|}{||z - z'||_E^{a_1 + \ldots + a_d}} < \infty, \text{ where } \nabla^a g(z) \equiv \frac{\partial^{a_1 + a_2 + \ldots + a_d}}{\partial z_1^{a_1} \ldots \partial z_d^{a_d}} g(z).$$

18
Assumption 4.3 (SA3) (i) \{Y_{1i}, Y_{2i}, W_i\}_{i=1}^n are i.i.d; (ii) \(\alpha_0 = (\theta_0, h_0) \in \mathcal{A} \equiv \Theta \times \mathcal{H}\); (iii) \(\Theta\) is compact with nonempty interior and \(\mathcal{H} = L_{C_1}^1(\mathcal{Z})\); (iv) \(K_n \to \infty\) and \(K_n/n \to 0\).

Lemma 4.2 (Consistency) Suppose Assumptions IS-A, SA2, ID, and SA3 hold. Suppose Condition 1 (Lipschitz conditions) in Appendix C holds. Then, \(\|\hat{\alpha}_n - \alpha_0\|_2 = o_p(1)\).

The proof can be found in the Appendix. This consistency result is obtained combining the identification condition and the uniform convergence of the sample criterion function. Next we consider the convergence rate of the sieve conditional ML estimator under a weaker metric\textsuperscript{14}. We obtain a convergence rate using Theorem 3.2 of Chen (2007) which is a version of Chen and Shen (1996) for i.i.d data. Now suppose that \(\mathcal{A} = \Theta \times \mathcal{H}\) is convex in \(\alpha_0\) such that \(\alpha_0 + \tau(\alpha - \alpha_0) \in \mathcal{A}\) for all small \(\tau \in [0, 1]\) and all fixed \(\alpha \in \mathcal{A}\). Suppose the pathwise derivative of \(l(\cdot)\) at the direction \([\alpha - \alpha_0]\) evaluated at \(\alpha_0\) by \(\frac{d(l(y_i|w_i,\alpha_0))}{d\alpha}[\alpha - \alpha_0] = \lim_{\tau \to 0} \frac{d(l(y_i|w_i,1-\tau\alpha+\tau\alpha))}{d\tau}(\theta - \theta_0) + \frac{d(l(y_i|w_i,\alpha_0))}{dh}[h - h_0]\). Define the \(L_2(P_0)\)-norm, \(\|\alpha - \alpha_0\|_2\), based on the pathwise derivative of \(l(\cdot)\) evaluated at \(\alpha_0\) as \(\|\alpha - \alpha_0\|_2^2 = E\left[\frac{d(l(Y_i|W_i,\alpha_0))}{d\alpha}[\alpha - \alpha_0]^2\right]\). This is the ML version of Ai and Chen (2003)'s \(L_2(P_0)\)-metric, which is a natural choice since it is the Fisher norm (see Wong and Severini (1991)) and the conditional ML version of the metric used by Wong and Severini (1991).

Proposition 4.2 Let \(\hat{\alpha}_n\) be the approximate sieve ML defined in (15). Suppose Assumptions IS-A, SA2, ID, and SA3 hold. Suppose Conditions 1 and 2-3 (Lipschitz conditions) in Appendix C hold. Then, we have \(\|\hat{\alpha}_n - \alpha_0\|_2 = O_p\left(\max \left\{\sqrt{K_n/n}, K_n^{1/2}\right\}\right)\). Moreover, with \(K_n = C\cdot n^{1/(2v_1+d_z)}\) and \(v_1/d_z > 1/2\), we have \(\|\hat{\alpha}_n - \alpha_0\|_2 = O_p\left(n^{-v_1/(2v_1+d_z)}\right)\).

The proof can be found in the Appendix.

4.2.2 Asymptotic Normality

In this section, we derive the \(\sqrt{n}\)-asymptotic normality of the structural parameters \(\hat{\theta}_n\). We show that even though we estimate the finite dimensional parameter \(\theta\) and the infinite dimensional parameter \(h\) simultaneously, the parametric rate inference for the finite dimensional parameter is still feasible. The following discussion and notation are based on Theorem 4.3 of Chen (2007), which is a simplified version of Shen (1997) and Chen and Shen (1998).

Suppose the functional of interest, \(f: \mathcal{A} \to \mathbb{R}\), is smooth in the sense that \(\frac{d(f(\alpha_0))}{d\alpha}[\alpha - \alpha_0] = \lim_{\tau \to 0} \frac{f(\alpha_0 + \tau(\alpha - \alpha_0)) - f(\alpha_0)}{\tau}\) is well-defined and \(\left\|\frac{d(f(\alpha_0))}{d\alpha}\right\| = \sup_{\alpha \in \mathcal{A}, \|\alpha - \alpha_0\|_2 > 0} \frac{|\frac{d(f(\alpha_0))}{d\alpha}[\alpha - \alpha_0]|}{\|\alpha - \alpha_0\|_2} < \infty\). Now let \(\mathcal{V}\) denote the closure of the linear span of \(\mathcal{A} - \alpha_0\) under the metric \(\|\alpha - \alpha_0\|_2\). Then, \((\mathcal{V}, \|\cdot\|_2)\) is a Hilbert space with inner product \(\langle v_1, v_2 \rangle = E\left[\frac{d(l(Y_i|W_i,\alpha_0))}{d\alpha}[v_1]\right] \left\|\frac{d(l(Y_i|W_i,\alpha_0))}{d\alpha}[v_2]\right\|^2\). Then, by the Riesz representation theorem, there exists \(v^* \in \mathcal{V}\) such that, for any \(\alpha \in \mathcal{A}\), \(\frac{d(f(\alpha_0))}{d\alpha}[\alpha - \alpha_0] = \langle \alpha - \alpha_0, v^* \rangle\) iff \(\left\|\frac{d(f(\alpha_0))}{d\alpha}\right\| < \infty\). In particular, we are interested in the linear functional \(f(\alpha) = \lambda^T\theta\)

\textsuperscript{14}Shen and Wong (1994), and Birgé and Massart (1998) derived the rates for general sieve M-estimation. Van de Geer (1993) and Wong and Shen (1995) derived the rates for sieve MLE for i.i.d data.
for any fixed and nonzero $\lambda \in \mathbb{R}^{d_{\theta}}$. Then, $f(\alpha) \equiv \lambda^t \theta$ is a linear function on $\mathbf{V}$. To estimate $f(\alpha) \equiv \lambda^t \theta$ at a $\sqrt{n}$ rate, $f(\alpha)$ has to be bounded (i.e. $\sup_{0 \neq \alpha - \alpha_0 \in \mathbf{V}} |f(\alpha) - f(\alpha_0)| / \|\alpha - \alpha_0\|_2 < \infty$) according to Van der Vaart (1991) and Shen (1997).

Now let $\mathcal{H} - h_0$ denote the closure of the linear span of $\mathcal{H} - h_0$. Define $b^*_j \in \mathbf{B} = \mathcal{H} - h_0$ for each component $\theta_j$ of $\theta$ such that $b^*_j = \arg\min_{b_j \in \mathbf{B}} E \left[ \left( \frac{d(Y|W,\alpha_0)}{dh} - \frac{d(Y|W,\alpha_0)}{dh} [b_j] \right)^2 \right]$. Now define $b^* = (b^*_1, \ldots, b^*_d)$, $\frac{d(Y|W,\alpha_0)}{dh}(b^*) = (\frac{d(Y|W,\alpha_0)}{dh}[b^*_1], \ldots, \frac{d(Y|W,\alpha_0)}{dh}[b^*_d])$, and $D_{b^*}(Y,W) = \frac{d(Y|W,\alpha_0)}{dh} - \frac{d(Y|W,\alpha_0)}{dh}[b^*]$. Here $D_{b^*}(Y,W)$ becomes the $1 \times d_{\theta}$ vector of efficient scores that are given by the mean projection residuals of the scores of the finite dimensional parameters $\frac{d(Y|W,\alpha_0)}{dh}$ on the scores associated with nuisance parameters $\frac{d(Y|W,\alpha_0)}{dh}[\cdot]$. For the inference we need additional regularity conditions that ensure the boundedness of $f(\alpha)$, the existence of well-behaving efficient scores for parametric components, and smoothness of each $b^*_j(z)$ that needs to be well-approximated in finite samples. We impose

**Assumption 4.4 (SA4)** (i) $\theta_0 \in \text{int}(\Theta)$; (ii) $E[D_{b^*}(Y,W)'D_{b^*}(Y,W)]$ is positive definite; (iii) each element $b^*_j(Z)$ belongs to the Hölder space $\Lambda_{C^j}^m(Z)$ with $m_j > d_z/2$.

Note that $\frac{d f(\alpha)}{d\alpha}[\alpha - \alpha_0] = (\theta - \theta_0)' \lambda$ and so $f(\alpha) - f(\alpha_0) - \frac{d f(\alpha)}{d\alpha}[\alpha - \alpha_0] = 0$. In addition, we can show for $f(\alpha) \equiv \lambda^t \theta$ with $\lambda \in \mathbb{R}^{d_{\theta}}, \lambda \neq 0$, $\sup_{\alpha \neq \alpha_0 \in \mathbf{V}} \frac{|f(\alpha) - f(\alpha_0)|^2}{\|\alpha - \alpha_0\|^2_2} = \lambda'(E[D_{b^*}(Y,W)'D_{b^*}(Y,W)])^{-1} \lambda$, which implies $f(\alpha) = \lambda^t \theta$ is bounded if and only if $E[D_{b^*}(Y,W)'D_{b^*}(Y,W)]$ is positive-definite, which is our Assumption SA4 (ii). For this case, there exists $v^* \in \mathbf{V}$ such that

$$f(\alpha) - f(\alpha_0) \equiv \lambda'(\theta - \theta_0) = (v^*, \alpha - \alpha_0) \text{ for all } \alpha \in \mathcal{A}$$

(16)

by the Riesz representation theorem. We find that $v^* \equiv (v^*_{\theta}, v^*_{\alpha}) \in \mathbf{V}$ satisfies (16) with $v^*_{\theta} = (E[D_{b^*}(Y,W)'D_{b^*}(Y,W)])^{-1} \lambda$ and $v^*_{\alpha} = -b^* \times v^*_{\theta}$. Moreover, Assumption SA4 (iii) ensures that $b^*$ can be well-approximated by sieves, so $v^*$ is also well-approximated by sieves. The following theorem states that we can achieve the $\sqrt{n}$-asymptotic normality for the finite dimensional parameters.

**Theorem 4.1** Suppose Assumptions IS-A, SA2, ID, SA3, and SA4 hold and suppose Conditions 1 and 2-3 (Lipschitz conditions) in the Appendix hold. Then, we have $\sqrt{n} (\hat{\theta}_n - \theta_0) \rightarrow_d N(0, \Omega_*)$ where $\Omega_* = \{E[D_{b^*}(Y,W)'D_{b^*}(Y,W)]\}^{-1}$.

The proof of this theorem can be found in the Appendix.

### 4.2.3 A Consistent Covariance Estimator

To do a statistical inference of the structural parameters based on Theorem 4.1, we need a consistent estimator of $\Omega_*$. For this we first need a consistent estimator of $b^*$. Similarly with Ai and Chen (2003), we can estimate $b^*_j$ by $\hat{b}^*_j$: $j = 1, \ldots, d_o$ as

$$\hat{b}^*_j = \arg\min_{b_j \in \mathcal{H}_n} \sum_{i=1}^n \left( \frac{d(\{y_i|w_i, \hat{\alpha}_n\})}{dh} - \frac{d(\{y_i|w_i, \hat{\alpha}_n\})}{dh}[b_j] \right)^2.$$ 

(17)
Note that for linear sieves, \( \hat{b}^*_j \) is easily obtained by regressing the derivatives of \( l(\cdot, \hat{\alpha}_n) \) with respect to \( \theta_j \) on the derivatives of \( l(\cdot, \alpha) \) with respect to \( h \). Finally, we estimate \( \Omega_\ast \) by \( \hat{\Omega}_\ast \equiv \left( \frac{1}{n} \sum_{i=1}^n D_{\theta_j}(y_i, w_i, \hat{\alpha}_n) D_{\theta_j}(y_i, w_i, \hat{\alpha}_n) \right)^{-1} \) where \( \hat{b}^* = (\hat{b}^*_1, \ldots, \hat{b}^*_d) \) and \( D_{\theta_j}(y, w, \alpha) = \frac{\partial l(y | w, \alpha)}{\partial \theta_j} \). We show that \( \hat{\Omega}_\ast \) is consistent under suitable regularity conditions.

Proposition 4.3 Suppose Assumptions IS-A, SA2, ID, SA3, and SA4 hold and suppose Conditions 1 and 2-3 (Lipschitz conditions) in the Appendix hold. Then, \( \hat{\Omega}_\ast = \Omega_\ast + o_p(1) \).

Note that for the consistency of our sieve ML estimator we need to promise \( K_n \) tends to infinity as the sample size gets large, i.e. we need to use more flexible specifications for \( h_n(\cdot) \) as \( n \) grows. However, in practice, one always does estimation with fixed \( K_n = K \). Even though the asymptotics will be different under two different scenarios (increasing or fixed \( K_n \)), the computed standard errors under two different scenarios can be numerically equivalent (see Ackerberg, Chen, and Hahn (2012)). Therefore, in practice we can ignore the semiparametric nature of our model and proceed both estimation and inference as if the parametric model is the true model.

4.3 Estimation of the Type Probability \( p \)

If it is of interest, we can distinguish between the population probability \( p \) of the strong type and the posterior belief \( p(Z) \) of Player 2, which is also the conditional probability of being strong type conditional on \( Z \). Identification of \( p \) becomes relevant if one wants to know the overall distribution of the strong type in the population. We can identify \( p \) from (2) as

\[
E[p(Z)] = \int p(z) f_Z(z) dz = \int (pf(s)(z) / f_Z(z)) f_Z(z) dz = p \int f(s)(z) dz = p. \tag{18}
\]

Then we can estimate \( p \) as \( \hat{p}_n = \frac{1}{n} \sum_{i=1}^n L(\hat{h}_n(Z_i)) = \frac{1}{n} \sum_{i=1}^n \exp(\hat{h}_n(z_i))/(1 + \exp(\hat{h}_n(z_i))) \) where \( \hat{h}_n(\cdot) \) is obtained from (15). Applying the mean value theorem with \( \hat{h}_n \) that lies between \( \hat{h}_n \) and \( h_0 \), we have

\[
\begin{align*}
\hat{p}_n - p_0 & = \frac{1}{n} \sum_{i=1}^n \left( L(\hat{h}_n(Z_i)) - L(h_0(Z_i)) \right) + \frac{1}{n} \sum_{i=1}^n \left( L(h_0(Z_i)) - E[L(h_0(Z_i))] \right) \\
& = \frac{1}{n} \sum_{i=1}^n L(\hat{h}_n(Z_i))(1 - L(\hat{h}_n(Z_i)))(\hat{h}_n(Z_i) - h_0(Z_i)) + \frac{1}{n} \sum_{i=1}^n \left( L(h_0(Z_i)) - E[L(h_0(Z_i))] \right) \\
& \leq (1/4)\|\hat{h}_n - h_0\|_\infty + o_p(1) \\
\end{align*}
\tag{19}
\]

where the second equality is obtained by \( L'(\cdot) = L(\cdot)(1 - L(\cdot)) \) and the last result holds since \( L(\cdot)(1 - L(\cdot)) \leq 1/4 \) and the second term in the RHS of (19) is \( o_p(1) \) by LLN noting \( |L(h_0(\cdot))| < 1 \) and \( \{Z_i\}_{i=1}^n \) are i.i.d.

Therefore, \( \hat{p}_n \) is consistent as long as \( \hat{h}_n \) is consistent. We can also derive the asymptotic distribution of \( \hat{p}_n \) following Chen, Linton, and van Keilegom (2003) since (18) can be written as a moment condition

\[
m(p, h) = p - L(h), E[m(p_0, h_0)] = 0
\]

and we have a pre-stage estimator \( \hat{h}_n \) from the sieve conditional ML. Let \( M(h) = \int_Z L(h) dF_Z \) (where \( F_Z \) denote the CDF of \( Z \)) and \( M_n(h) = \frac{1}{n} \sum_{i=1}^n L(h(Z_i)) \). We state the asymptotic normality of \( \hat{p}_n \) as follows.
Proposition 4.4 Suppose (i) \( \| \hat{h}_n - h_0 \|_\infty = o_p(n^{-1/4}) \) and (ii)
\[
\sqrt{n} \left( \int_Z L(h_0)(1 - L(h_0)) (\hat{h}_n - h_0) dF_Z + M_n(h_0) - M(h_0) \right) \to_d N(0, V_p).
\]
Then, \( \sqrt{n} (\hat{p}_n - p_0) \to_d N(0, V_p) \).

Note that the first term in the LHS of (20) appears due to the fact that we use a pre-stage estimator \( \hat{h}_n \). Note that we can satisfy the condition (i) by combining the result in Proposition 4.2 and the fact \( \| h - h_0 \|_\infty \leq \| h - h_0 \|_2^{2\nu/(2\nu_1 + d)} \) (Lemma 2 in Chen and Shen (1998)). In the Appendix, we also discuss how the condition (ii) (also the condition (iv) in Proposition 4.5 below) can be satisfied for the sieve conditional ML estimator under Assumptions in Theorem 4.1.

Finally, we note that even if an explicit form of \( V_p \) can be derived, a feasible estimator of \( V_p \) may be difficult to calculate. Alternatively, we can use a nonparametric bootstrap. The following proposition shows that we can approximate the distribution of \( \sqrt{n} (\hat{p}_n - p_0) \) using a bootstrap. We use “*” to denote the bootstrap counterpart of the original sample \( \{Z_i\}_{i=1}^n \). We let \( M^*_n(h) = \frac{1}{n} \sum_{i=1}^n L(h(Z^*_i)) \).

Proposition 4.5 Suppose (i) with \( P^* \)-probability tending to one, \( \hat{h}^*_n \in H \), and \( \| \hat{h}^*_n - \hat{h}_n \|_\infty = o_P^*(n^{-1/4}) \); (ii) \( \| \hat{h}_n - h_0 \|_\infty = o(n^{-1/4}) \) a.s.; for any positives sequence \( \delta_n \) tending to zero,
\[
(iii) \sup_{\| h - h_0 \|_\infty \leq \delta_n} | M_n(h) - M(h) - M_n(h_0) + M(h_0) | = o(n^{-1/2}) \text{ a.s.} \\
(iv) \sqrt{n} \left( \int_Z L(h_n)(1 - L(h_n)) (\hat{h}^*_n - \hat{h}_n) dF_Z + M^*_n(\hat{h}_n) - M_n(\hat{h}_n) \right) \to N(0, V_p) + o_P^*(1).
\]
Then, \( \sqrt{n} (\hat{p}_n^* - \hat{p}_n) \) converges in distribution to \( N(0, V_p) \) in \( P^* \)-probability.

5 Monte Carlo Simulation

In this section we investigate the finite sample performance of our proposed estimator. We conduct a small-scale Monte Carlo experiment and show that the conditional ML estimator can recover the true parameters reasonably well. We first generate payoffs and type relevant variables and then generate equilibrium actions of players using the game structure (Figure 2) as below.

In the Monte Carlo design we let \( X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2}, Z_1, \) and \( Z_2 \sim N(0, 1) \), respectively and let \( X_1 = [X_{1,1}, X_{1,2}], X_2 = [X_{2,1}, X_{2,2}], \) and \( Z = [Z_1, Z_2] \). In all designs below \( X_1 \) and \( X_2 \) are jointly independent each other. The payoffs shocks \((\varepsilon_1, \varepsilon_2)\) follow independent standard normal distributions. For the true payoffs parameters we take \( \beta_1 = (1, -1) \) and \( \beta_2 = (1, -1) \) and for the strategic interaction parameters we take \( (\phi_{1s}, \phi_{1w}, \phi_{2s}, \phi_{2w}) = (1, 1, 1, 1) \). Here we can estimate these interaction parameters separately for each type because there is no intercept in Player’s payoffs (this is a normalization).

We experiment with a few different settings. In Design [1] we let \( X_1, X_2, \) and \( Z \neq X_1 \) be all mutually independent and model the probability of being a strong type for Player 1 with her characteristics \( Z \) as \( p(Z) = \exp(\pi_0 + Z \pi_1)/(1 + \exp(\pi_0 + Z \pi_1)) \) and treat this \( p(Z) \) as the model primitive. We take \( (\pi_0, \pi_1) = (0.2, 1, 1) \) for the true values. In Design [2] we let \( Z = X_1 \) and use the same specification of \( p(Z) \) in Design [1], so the same variables \( X_1 \) both affect the payoffs and inform about the type of Player 1. In Design [3] we assume there is no public signal, so \( p(Z) = p \)
and we take \( p = 0.2 \). In Design [4] we consider a semiparametric model for which we pretend “we do not know” the functional form of \( p(Z) \). We use the following as the true \( p(Z) \):

\[
P(Z) = \exp(h(Z))/(1 + \exp(h(Z))) \quad \text{with} \quad h(Z) = 0.2Z + \log(Z^2/2 + 1)/5
\]

where \( Z \sim N(0, 1) \). Given these model primitives for each design we generate \( T = 200 \) and 500 observations by first simulating the model choice probabilities at the true parameter value and generating the observed actions.

In our sieve estimation of Design [4] we approximate the “unknown” \( h(Z) \) using the Hermite polynomials approximation (e.g. Newey and Powell 2003) as

\[
h(Z) \approx \sum_{k=1}^{K} a_k \exp(-Z^2)Z^{k-1}
\]

and we take \( K = 3 \) or 4. Table 1 and 2 summarize estimation results based on these MC designs and we calculate the bias and the RMSE (root mean squared error), averaged across 50 repetitions of each design. For the estimates of \( p(Z) \) we also average across the realized values of \( Z \) to obtain the bias and the RMSE.

The results show that the game parameters are reasonably well estimated from the conditional ML. The performance of the estimator in terms of RMSE generally improves as the sample size grows. This is true whether or not there exist public signals: Design [1]&[2] vs Design [3]. By comparing Designs [1] and [2] we also find that the estimator for Player 1’s game parameters in Design [1] overall performs slightly better than in Design [2], which suggests that the exclusion restriction helps for identification because variables that vary payoffs are different from those that vary public signals.

In the semiparametric estimation of Design [4], where we do not know the true functional form of \( p(Z) \), the estimator performs reasonably well, suggesting our sieve approach can suitably approximate the true function. Comparing the semiparametric estimation results with \( K = 3 \) and \( K = 4 \) we find the estimates are similar in both bias and RMSE, suggesting the third order term in the sieve approximation with \( K = 4 \) is less important. Although it is a limited evidence in any nature, these Monte Carlo experiments overall suggest that the proposed conditional ML estimator works reasonably well.

### Table 1. Monte Carlo Experiment with T=200: Parametric Models

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_{1,1} )</td>
<td>Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td>-0.161</td>
<td>0.0461</td>
<td>0.0022</td>
</tr>
<tr>
<td>-0.0069</td>
<td>0.0614</td>
<td>0.0720</td>
</tr>
<tr>
<td>( \beta_{2,1} )</td>
<td>0.0258</td>
<td>0.0293</td>
</tr>
<tr>
<td>( \beta_{2,2} )</td>
<td>-0.0272</td>
<td>0.0308</td>
</tr>
<tr>
<td>( \phi_{1s} )</td>
<td>0.0649</td>
<td>0.0844</td>
</tr>
<tr>
<td>( \phi_{1w} )</td>
<td>0.0408</td>
<td>0.0727</td>
</tr>
<tr>
<td>( \phi_{2s} )</td>
<td>-0.0098</td>
<td>0.0215</td>
</tr>
<tr>
<td>( \phi_{2w} )</td>
<td>-0.0182</td>
<td>0.0249</td>
</tr>
<tr>
<td>( \pi_0 )</td>
<td>0.0023</td>
<td>0.0140</td>
</tr>
<tr>
<td>( \pi_1 )</td>
<td>0.0143</td>
<td>0.0176</td>
</tr>
<tr>
<td>( \pi_2 )</td>
<td>0.0068</td>
<td>0.0124</td>
</tr>
<tr>
<td>( p )</td>
<td>-0.0029</td>
<td>0.0032</td>
</tr>
</tbody>
</table>
### Table 2. Monte Carlo Experiment with T=500: Parametric and Semiparametric Models

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Signal</td>
<td>Public Signal</td>
<td>No Public Signal</td>
<td>Unknown $p(Z)$</td>
</tr>
<tr>
<td>$Z \neq X_1$</td>
<td>$Z = X_1$</td>
<td>$K = 3$</td>
<td>$K = 4$</td>
</tr>
<tr>
<td>$eta_{1,1}$</td>
<td>.0128</td>
<td>.0489</td>
<td>Bias RMSE</td>
</tr>
<tr>
<td></td>
<td>-.0128</td>
<td>.0489</td>
<td>Bias RMSE</td>
</tr>
<tr>
<td>$eta_{1,2}$</td>
<td>.0183</td>
<td>.0441</td>
<td>.0193</td>
</tr>
<tr>
<td></td>
<td>.0294</td>
<td>.0330</td>
<td>.0316</td>
</tr>
<tr>
<td></td>
<td>-.0310</td>
<td>.0346</td>
<td>-.0316</td>
</tr>
<tr>
<td></td>
<td>.0582</td>
<td>.0650</td>
<td>.0794</td>
</tr>
<tr>
<td></td>
<td>.0328</td>
<td>.0591</td>
<td>.0643</td>
</tr>
<tr>
<td></td>
<td>-.0029</td>
<td>.0206</td>
<td>-.0100</td>
</tr>
<tr>
<td></td>
<td>-.0127</td>
<td>.0271</td>
<td>-.0051</td>
</tr>
<tr>
<td></td>
<td>.0021</td>
<td>.0089</td>
<td>-.0002</td>
</tr>
<tr>
<td></td>
<td>.0169</td>
<td>.0179</td>
<td>.0162</td>
</tr>
<tr>
<td></td>
<td>.0170</td>
<td>.0188</td>
<td>.0080</td>
</tr>
<tr>
<td>$p$</td>
<td>-.0027</td>
<td>.0031</td>
<td>Bias RMSE</td>
</tr>
<tr>
<td>$p(Z)$</td>
<td>-.0118</td>
<td>.0232</td>
<td>-.0098</td>
</tr>
</tbody>
</table>

### 6 Conclusion

This paper develops an econometric modeling of a signaling game with two players where one player, *informed player*, has private information on its type. In particular, we provide an estimation strategy that identifies the payoffs structure and the conditional probabilities of being a particular type from the observed actions. Multiplicity of equilibria arises in this game and we show that the uniqueness of equilibrium, given any realization of payoffs, can be achieved using the equilibrium refinement of Cho and Kreps (1987). This uniqueness result enables us to derive well-defined model choice probabilities.

Our model also accommodates public signals about the type of the informed player. The *uninformed player* will use this information to update her belief on types after observing an action of the *informed player*. Since the distribution of these public signals is nonparametrically specified, we estimate the model using a sieve conditional MLE where the infinite dimensional parameters are approximated by sieves. We obtain the consistency and the root $n$-asymptotic normality of the structural parameter estimates. We also run a small-scale Monte Carlo experiment to illustrate the finite sample performance of the proposed estimator.
Appendix

A Equilibrium Refinement and Uniqueness of Equilibrium (Proof of Theorem 3.2)

We show that in all regions where multiple PBE exist (see Figure 3) only one unique PBE survives the Intuitive Criterion of Cho and Kreps (1987) (see Figure 4).

[1] For $M_1 = \{(\varepsilon_1, \varepsilon_2) | \varepsilon_1 > X'_1 \beta_1 + \phi_{1w} \}$ and $X'_2 \beta_2 + (\phi_{2s} + \phi_{2w})p(\cdot) - \phi_{2w} < \varepsilon_2 < X'_2 \beta_2 + \phi_{2s}$, we show the pooling with $B$ cannot survive the Intuitive Criterion while the pooling with $Q$ does.

- Pooling with $(A_{1t_s}, A_{1t_w}) = (B, B)$ and $A_{2|B} = F$:
  Note $u_1(t_s; B, F) > u_1(t_s; Q, NF)$ since $X'_1 \beta_1 - \varepsilon_1 - \phi_{1w} < 0$ under $M_1$ and note $u_1(t_s; B, F) > u_1(t_s; Q, F)$ since $\phi_{1w} > 0$. This means that the strong type has no incentive to deviate in any case. Thus, Player 2 will assign $\mu_2(t_1 = t_s | Q) = 0$ and hence Player 2 will choose $F$ after observing the deviation play $Q$ under the region $M_1$. Now we need to check whether the weak type is better off by deviation under this situation. Note $u_1(t_w; B, F) = -\phi_{1w} < u_1(t_w; Q, F) = 0$ and thus the weak type will deviate for sure. Therefore, this equilibrium fails the Intuitive Criterion.

- Pooling with $(A_{1t_s}, A_{1t_w}) = (Q, Q)$ and $A_{2|Q} = F$:
  Note $u_1(t_w; Q, F) > u_1(t_w; B, NF)$ since $X'_1 \beta_1 - \varepsilon_1 - \phi_{1w} < 0$ under $M_1$ and note $u_1(t_w; Q, F) > u_1(t_w; B, F)$ since $\phi_{1w} > 0$. This means that the weak type has no incentive to deviate in any case. Thus, Player 2 will assign $\mu_2(t_1 = t_s | B) = 1$ and hence Player 2 will choose $NF$ after observing the deviation play $B$ under the region $M_1$. Now we need to check whether the strong type is better off by deviation under this situation. Note $u_1(t_s; Q, F) = -\phi_{1s} > u_1(t_s; B, NF) = X'_1 \beta_1 - \varepsilon_1$ under $S_4$ and thus the strong type will not deviate. Therefore, this equilibrium survives the Intuitive Criterion.

[2] For $M_2 = \{(\varepsilon_1, \varepsilon_2) | \varepsilon_1 < X'_1 \beta_1 - \phi_{1s} \}$ and $X'_2 \beta_2 - \phi_{2w} < \varepsilon_2 < X'_2 \beta_2 + (\phi_{2s} + \phi_{2w})p(\cdot) - \phi_{2w}$, we show the pooling with $Q$ cannot survive the Intuitive Criterion while the pooling with $B$ does.

- Pooling with $(A_{1t_s}, A_{1t_w}) = (Q, Q)$ and $A_{2|Q} = NF$:
  Note $u_1(t_w; Q, NF) > u_1(t_w; B, F)$ since $X'_1 \beta_1 - \varepsilon_1 + \phi_{1w} > 0$ under $M_2$ and note $u_1(t_w; Q, NF) > u_1(t_w; B, NF)$ since $\phi_{1w} > 0$. This means that the weak type has no incentive to deviate in any case. Thus, Player 2 will assign $\mu_2(t_1 = t_s | B) = 1$ and hence Player 2 will choose $NF$ after observing the deviation play $B$ under the region $M_2$. Now we need to check whether the strong type is better off by deviation under this situation. Note $u_1(t_s; Q, NF) = X'_1 \beta_1 - \varepsilon_1 - \phi_{1s} < u_1(t_s; B, NF) = X'_1 \beta_1 - \varepsilon_1$ since $\phi_{1s} > 0$. Thus, the strong type will deviate for sure. Therefore, this equilibrium fails the Intuitive Criterion.

- Pooling with $(A_{1t_s}, A_{1t_w}) = (B, B)$ and $A_{2|B} = NF$:
  Note $u_1(t_s; B, NF) > u_1(t_s; Q, NF)$ since $\phi_{1s} > 0$ and note $u_1(t_s; B, NF) > u_1(t_s; Q, F)$ under $M_2$. This means that the strong type has no incentive to deviate in any case. Thus, Player 2 will assign $\mu_2(t_1 = t_s | Q) = 0$ and hence Player 2 will choose $F$ after observing the deviation play $Q$ under the region $M_2$. Now we need to check whether the weak type is better
off by deviation under this situation. Note \( u_1(t_w;B,NF) = X'_1\beta_1 - \varepsilon - \phi_{1w} > u_1(t_w;Q,F) = 0 \) under \( S_5 \). Thus, the weak type will not deviate. Therefore, this equilibrium survives the Intuitive Criterion.

[3] The separating equilibrium \((A_{1t}, A_{1w}) = (B, Q)\) with \((A_{2|B} = NF, A_{2|Q} = F)\) under the region \( S_3 \) cannot fail the Intuitive Criterion because none of Player 1 wants to deviate regardless of Player 2’s action.

B Conditional Probabilities of Observed Outcomes

In the Supplementary Appendix, we derive all possible PBE of the signaling game. Then, applying the refinement of Cho and Kreps (1987) we achieve the uniqueness of equilibrium. This enables us to derive the conditional probabilities of the observed outcomes given the parameterization of the signaling game. Here we summarize the conditional probabilities in the game with IS-A and SA2. For the game with IS and SA, we obtain the same forms of conditional probabilities by replacing \( p(Z) \) with \( p, \) so \( W \) with \( X \). Detailed derivations of these probabilities are provided in the Supplementary Appendix.

For the conditional ML estimation using these model probabilities we need to normalize either \((\beta_{1,0} = 0, \beta_{2,0} = 0)\) or \((\phi_{1s} = \phi_{1w}, \phi_{2s} = \phi_{2w})\) below.

1 \((Y_1, Y_2) = (1, 1)\): It occurs under \( S_5 \) with probability one (pooling), under \( S_3 \cup S_7 \) with probability \( p(Z) \) (separating), and under \( S_2 \cup S_6 \) (semi-separating). Combining the measures of these regions, we obtain

\[
\text{Pr}(Y_1 = 1, Y_2 = 1|W; \alpha) = G_1(X'_1\beta_1 - \phi_{1w}) (G_2(X'_2\beta_2 + p(Z)(\phi_{2s} + \phi_{2w}) - \phi_{2w}) - G_2(X'_2\beta_2 - \phi_{2w})) + p(G_1(X'_1\beta_1 + \phi_{1s}) - G_1(X'_1\beta_1 - \phi_{1w})) (G_2(X'_2\beta_2 + \phi_{2s}) - G_2(X'_2\beta_2 - \phi_{2w})) + pG_2(X'_2\beta_2 - \phi_{2w}) + \int_0^1 \left( p(\cdot) + (1 - p(\cdot))\mu \right) g_2(X'_2\beta_2 + \left( \frac{p(\cdot)}{p(\cdot) + \mu(1-p(\cdot))} \right) (\phi_{2s} + \phi_{2w}) - \phi_{2w}) \frac{p(\cdot)(1-p(\cdot))(\phi_{2s} + \phi_{2w})}{(p(\cdot) + (1-p(\cdot))\mu)^2} d\mu \times \int_0^1 \sigma_2 g_1(X'_1\beta_1 - \phi_{1w}/\sigma_2)(\phi_{2s} + \phi_{2w}) - \phi_{2w}) \frac{p(\cdot)(1-p(\cdot))(\phi_{2s} + \phi_{2w})}{(1-\mu p(\cdot))^2} d\mu \times \int_0^1 \sigma_2 g_1(X'_1\beta_1 + \phi_{1s}/(1 - \sigma_2)) \phi_{1s}/(1 - \sigma_2)^2 d\sigma_2
\]

2 \((Y_1, Y_2) = (1, 0)\): It occurs under \( S_1 \) with probability \( p(Z) \) (separating) and under \( S_2 \cup S_6 \) (semi-separating). We obtain

\[
\text{Pr}(Y_1 = 1, Y_2 = 0|W; \alpha) = p(\cdot)(1 - G_2(X'_2\beta_2 + \phi_{2s})) + \int_0^1 \left( p(\cdot) + (1 - p(\cdot))\mu \right) g_2(X'_2\beta_2 + \left( \frac{p(\cdot)}{p(\cdot) + \mu(1-p(\cdot))} \right) (\phi_{2s} + \phi_{2w}) - \phi_{2w}) \frac{p(\cdot)(1-p(\cdot))(\phi_{2s} + \phi_{2w})}{(p(\cdot) + (1-p(\cdot))\mu)^2} d\mu \times \int_0^1 (1 - \sigma_2) g_1(X'_1\beta_1 - \phi_{1w}/\sigma_2)(\phi_{2s} + \phi_{2w}) - \phi_{2w}) \frac{p(\cdot)(1-p(\cdot))(\phi_{2s} + \phi_{2w})}{(1-\mu p(\cdot))^2} d\mu \times \int_0^1 (1 - \sigma_2) g_1(X'_1\beta_1 + \phi_{1s}/(1 - \sigma_2)) \phi_{1s}/(1 - \sigma_2)^2 d\sigma_2
\]
3 \((Y_1, Y_2) = (0, 1)\): It occurs under \(S_7\) with probability \(1 - p(Z)\) (separating), and under \(S_2 \cup S_6\) (semi-separating). We obtain
\[
\Pr(Y_1 = 0, Y_2 = 1|W, \alpha) = (1 - p(\cdot)) G_2(X'_2 \beta_2 + \phi_2w) + \int_0^1 (1 - p(\cdot))(1 - \mu) g_2(X'_2 \beta_2 + \phi_2w) \frac{p(\cdot)(1 - p(\cdot))(\phi_2s + \phi_2w)}{(p(\cdot) + \mu(1 - p(\cdot)))^2} d\mu \\
\times \int_0^1 2\sigma_2 g_1(X'_1 \beta_1 - \phi_1w/\sigma_2) 1w/\sigma_2^2 d\sigma_2
\]
and thus proves the claim.

4 \((Y_1, Y_2) = (0, 0)\): It occurs under \(S_4\) with probability one (pooling), under \(S_1 \cup S_3\) with probability \(1 - p(Z)\) (separating), and under \(S_2 \cup S_6\) (semi-separating). We obtain
\[
\Pr(Y_1 = 0, Y_2 = 0|W, \alpha) = (1 - G_1(X'_1 \beta_1 + \phi_1w)) (G_2(X'_2 \beta_2 + \phi_2w) - G_2(X'_2 \beta_2 + p(\cdot)(\phi_2s + \phi_2w))) + (1 - p(\cdot)) (1 - G_2(X'_2 \beta_2 + \phi_2w)) + (1 - p(\cdot))(G_1(X'_1 \beta_1 + \phi_1w) - G_1(X'_1 \beta_1 - \phi_1w)) (G_2(X'_2 \beta_2 + \phi_2s) - G_2(X'_2 \beta_2 - \phi_2w)) + \int_0^1 (1 - p(\cdot))(1 - \mu) g_2(X'_2 \beta_2 + \phi_2w) \frac{p(\cdot)(1 - p(\cdot))(\phi_2s + \phi_2w)}{(p(\cdot) + \mu(1 - p(\cdot)))^2} d\mu \\
\times \int_0^1 (1 - \sigma_2) g_1(X'_1 \beta_1 - \phi_1w/\sigma_2) 1w/\sigma_2^2 d\sigma_2 + \int_0^1 (1 - \mu p(\cdot)) g_2(X'_2 \beta_2 + \phi_2s) \frac{p(\cdot)(1 - p(\cdot))(\phi_2s + \phi_2w)}{(1 - \mu p(\cdot))^2} d\mu \\
\times \int_0^1 (1 - \sigma_2) g_1(X'_1 \beta_1 + \phi_1w/(1 - \sigma_2)) \phi_1w/(1 - \sigma_2)^2 d\sigma_2.
\]

C Large Sample Theories for the Sieve Conditional ML

C.1 Identification (Proof of Lemma 4.1)

Let \(\mathcal{L}(Y, W, \alpha) = \exp(l(Y|W, \alpha))\) and note
\[
\frac{\mathcal{L}(Y, W, \alpha)}{\mathcal{L}(Y, W, 0)} \in \left\{ \frac{\Pr(Y = y|W, \alpha)}{\Pr(Y = y|W, 0)} : y \in \{(1, 1), (1, 0), (0, 1), (0, 0)\} \right\}
\]
where \(\Pr(Y|W, \alpha)\) denotes the conditional probability of \(Y\) given \(W = X \cup Z\) when the parameter equals to \(\alpha\). Applying Jensen’s inequality, we have
\[
-\ln \left\{ E \left[ \mathcal{L}(Y, W, \alpha) / \mathcal{L}(Y, W, 0) \right] \right\} < -E \left[ \ln \left\{ \mathcal{L}(Y, W, \alpha) / \mathcal{L}(Y, W, 0) \right\} \right]
\]
(21)
since \(\mathcal{L}(Y, W, \alpha) / \mathcal{L}(Y, W, 0)\) is always positive and not constant whenever \(\alpha \neq \alpha_0\) by Assumption ID. We also have \(\Pr(\mathcal{L}(Y, W, \alpha) / \mathcal{L}(Y, W, 0)) = \Pr(Y = y|W, \alpha) / \Pr(Y = y|W, 0))\) = \(\Pr(Y = y|W, \alpha)\) under Assumptions IS-A and SA2. It follows that
\[
E \left[ \frac{\mathcal{L}(Y, W, \alpha)}{\mathcal{L}(Y, W, 0)} \right] = \int \left\{ \sum_y \Pr(Y = y|W, \alpha) \cdot \Pr(Y = y|W, 0) \right\} f_W(W) dW
\]
(22)
where the last equality holds since \(\sum_y \Pr(Y = y|W, \alpha) = 1\) for all \(\alpha \in \mathcal{A}\). Therefore, combining (21) and (22), we conclude that for all \(\alpha \neq \alpha_0 \in \mathcal{A}\), \(0 < E \left[ \ln \mathcal{L}(Y, W, 0) \right] - E \left[ \ln \mathcal{L}(Y, W, \alpha) \right]\). This implies \(Q(\alpha_0) > Q(\alpha)\) and thus proves the claim.
C.2 Consistency

To prove the consistency, we need additional assumptions. For ease of notation we use the normalization \( \phi_1 \equiv \phi_{1s} = \phi_{1w} \) and \( \phi_2 \equiv \phi_{2s} = \phi_{2w} \).

**Condition 1 (Lipschitz Condition)**

(i) For some convex combination of \( \alpha_1, \alpha_2 \in A_n \), there exist functions \( M_{ij}^{(\phi_1)}(\cdot) \), \( M_{ij}^{(\phi_2)}(\cdot) \), \( M_{ij}^{(\beta_1)}(\cdot) \), \( M_{ij}^{(\beta_2)}(\cdot) \), and \( M_{ij}^{(h)}(\cdot) \) such that

\[
\frac{dL_{ij}(W, \alpha)}{d\alpha} [\alpha_1 - \alpha_2] = M_{ij}^{(\phi_1)}(W, \overline{\alpha})(\phi_{11} - \phi_{12}) + M_{ij}^{(\phi_2)}(W, \overline{\alpha})(\phi_{21} - \phi_{22}) + M_{ij}^{(\beta_1)}(W, \overline{\alpha})X_1^i(\beta_{11} - \beta_{12}) \\
+ M_{ij}^{(\beta_2)}(W, \overline{\alpha})X_2^i(\beta_{21} - \beta_{22}) + M_{ij}^{(h)}(W, \overline{\alpha})(h_1 - h_2)
\]

for all \( i, j = 0, 1 \);

(ii) \( \sup_{\alpha \in A_n} |M_{ij}^{(t)}(W, \alpha)| \leq C_t(W) < \infty \) and \( E[C_t^2(W)] < \infty \), \( \forall t \in \{\phi_1, \phi_2, \beta_1, \beta_2, h\} \), \( \forall i, j = 0, 1 \), \( n \geq \exists N \).

This condition is not difficult to verify as we discuss in Section E. We reproduce Theorem 3.1 of Chen (2007) in terms of the sieve conditional ML estimator defined in (15).

**Theorem C.1 (Theorem 3.1 in Chen (2007))**

Suppose \((C1)\) \( Q(\alpha) \) is uniquely maximized on \( A \) at \( \alpha_0 \in A \), and \( Q(\alpha_0) > -\infty \);

\((C2)\) \( A_n \subset A_{n+1} \subset A \) for all \( n \geq 1 \), and for any \( \alpha \in A \) there exists \( \Pi n \alpha \in A_n \) such that \( \| \Pi n \alpha - \alpha \|_s \to 0 \) as \( n \to \infty \);

\((C3)\) The criterion function, \( Q(\alpha) \), is continuous in \( \alpha \in A \) with respect to \( \| \|_s \);

\((C4)\) The sieve spaces, \( A_n \), are compact under \( \| \|_s \);

\((C5)\) \( \lim_{n \to \infty} \sup_{\alpha \in A_n} |\hat{Q}_n(\alpha) - Q(\alpha)| = 0 \) holds.

Let \( \hat{\alpha}_n \) be the approximate sieve ML estimator defined by (15), then \( \| \hat{\alpha}_n - \alpha \|_s = o_p(1) \).

Below we verify Conditions \((C1)-(C5)\) and thus prove the consistency. Note that Condition \((C1)\) is satisfied by Lemma 4.1. Condition \((C2)\) holds for the sieve space \( A_n = \Theta \times H_n \) with \( H_n \) defined in (13) (see Section 5.3.2 of Timan (1963)). Condition \((C3)\) is satisfied since each \( L_{ij} \), \( i, j = 1, 0 \) is (pointwise) Lipschitz continuous by Condition 1. Condition \((C4)\) holds for the sieve space (13). Now let \( F_n = \{l(y|w, \theta, h) : (\theta, h) \in A_n\} \) denote the class of measurable functions indexed by \( (\theta, h) \). Condition \((C5)\) will be satisfied, for example, if \( F_n \) is P-Glivenko-Cantelli or P-Donsker (see e.g. van der Vaart and Wellner (1996)). The following lemma establishes this uniform convergence result.

**Lemma C.1 (Uniform convergence)**

Suppose Assumptions \( ID \) and \( SA3 \) hold. Then, for \( \hat{Q}_n(\alpha) \) and \( Q(\alpha) \) defined in (14) and (4) and for \( A_n = \Theta \times H_n \) with \( H_n \) defined in (13), we have \( \lim_{n \to \infty} \sup_{\alpha \in A_n} |\hat{Q}_n(\alpha) - Q(\alpha)| = 0 \).

**Proof.** We prove this lemma by showing that all the conditions (i), (ii), and (iii) of Lemma A2 in Newey and Powell (2003) are satisfied. The condition (i) is satisfied for \( A_n = \Theta \times H_n \) with \( H_n \)
defined in (13) and for the metric \( \| \cdot \|_a \). The condition (ii) is satisfied if \( E[|l(Y_i|W_i, \theta, h(Z_i))|] < \infty \), for all \((\theta, h) \in \mathcal{A}_n\) by the law of large numbers. Note that this condition is satisfied since \( \mathcal{L}_{ij}(W, \theta, h) \), \( \forall i, j = 0, 1 \) is uniformly bounded between 0 and 1 over \( \mathcal{A}_n \) and because under Assumptions \textbf{SA2}, \( \Pr (\mathcal{L}_{ij}(W, \theta, h) = 0 \text{ or } \mathcal{L}_{ij}(W, \theta, h) = 1) = 0 \) for \( \forall i, j = 0, 1 \) and all \((\theta, h) \in \mathcal{A}_n\), where \( \Pr (\cdot) \) is a probability measure over \( W \). Therefore, the condition (ii) holds. The condition (iii) is also satisfied since \( \hat{Q}_n(\alpha) \) is Lipschitz with respect to \( \alpha \) by Condition 1. This completes the proof. ■

Therefore, under Assumptions \textbf{ID}, \textbf{SA2}, and \textbf{SA3} and Condition 1, all the conditions in Theorem C.1 are satisfied for \( \hat{\alpha}_n \). This establishes the consistency result of the sieve ML estimator.

### C.3 Convergence Rate

We introduce additional notation. For any \( \alpha_1, \alpha_2, \alpha_3 \in \mathcal{A} \), the pathwise first derivatives are defined as \( \frac{dl(y|w, \alpha_0)}{d\alpha} [\alpha_1 - \alpha_2] \equiv \frac{dl(y|w, \alpha_0)}{d\alpha} [\alpha_1 - \alpha_0] - \frac{dl(y|w, \alpha_0)}{d\alpha} [\alpha_2 - \alpha_0] \equiv \lim_{\tau \to 0} \frac{dl(y|w, \alpha_3 + \tau(\alpha_1 - \alpha_0))}{d\tau} \). The pathwise second derivatives are defined in a similar way such that \( \frac{d^2l(y|w, \alpha_0)}{d\alpha^2} [\alpha_1 - \alpha_0, \alpha_2 - \alpha_0] \equiv \lim_{\tau \to 0} d\left( \frac{dl(y|w, \alpha_3 + \tau(\alpha_1 - \alpha_0))}{d\alpha} \right) [\alpha_1 - \alpha_0]/d\tau \). In particular, the pathwise second derivative at the direction \( \alpha - \alpha_0 \) evaluated at \( \alpha_0 \) is denoted by \( \frac{d^2l(y|w, \alpha_0)}{d\alpha^2} [\alpha - \alpha_0, \alpha - \alpha_0] \equiv \lim_{\tau \to 0} \frac{dl(y|w, \alpha_0 + \tau(\alpha_1 - \alpha_0))}{d\tau} \). Note that we have the information matrix equality as \( E \left[ \frac{dl(Y|W, \alpha_0)}{d\alpha} [\alpha_1 - \alpha_0] \frac{dl(Y|W, \alpha_0)}{d\alpha} [\alpha_2 - \alpha_0] \right] = -E \left[ \frac{d^2l(Y|W, \alpha_0)}{d\alpha^2} [\alpha_1 - \alpha_0, \alpha_2 - \alpha_0] \right] \) by Wong and Severini (1991).

Here we derive a convergence rate of the sieve ML estimator defined in (15) under the metric \( \| \cdot \|_2 \). We use Theorem 3.2 of Chen (2007), which is a version of Chen and Shen (1996) for i.i.d data. Before stating Theorem 3.2 of Chen (2007) and proving Proposition 4.2, we introduce additional notation.

Let \( K(\alpha_0, \alpha) = n^{-1} \sum_{i=1}^{n} E[|l(Y_i|W_i, \alpha_0) - l(Y_i|W_i, \alpha)|] \) denote the Kullback-Leibler information (divergence measure) on \( n \) observations. Let \( \| \cdot \|_a \) be a metric on \( \mathcal{A} \) which is equivalent to \( K(\cdot, \cdot)^{1/2} \). The equivalence means there exist constants \( C_1 \) and \( C_2 > 0 \) such that \( C_1 K(\alpha_0, \alpha)^{1/2} \leq \| \alpha_0 - \alpha \|_a \leq C_2 K(\alpha_0, \alpha)^{1/2} \) for all \( \alpha \in \mathcal{A} \). It is well known in the literature that the convergence rate depends on two factors. One is how fast the sieve approximation error rate, \( \| \alpha_0 - \Pi_n \alpha_0 \|_a \), goes to zero. The other is the complexity of the sieve space \( \mathcal{A}_n \). Let \( L_r(P_0), \ r \in [1, \infty) \) denote the space of real-valued random variables with finite \( r \)-th moments and \( \| \cdot \|_{L_r} \) denote the \( L_r(P_0) \)-norm. Let \( \mathcal{F}_n = \{ g(\alpha, \cdot) : \alpha \in \mathcal{A}_n \} \) be a class of real-valued, \( L_r(P_0) \)-measurable functions indexed by \( \alpha \in \mathcal{A}_n \). We let \( N(\epsilon, \mathcal{F}_n, \| \cdot \|_{L_r}) \) denote the covering numbers without bracketing, which is the minimal number of \( \epsilon \)-balls that covers \( \mathcal{F}_n \). We often use the notion of \( L_r(P_0) \)-metric entropy, \( H(\epsilon, \mathcal{F}_n, \| \cdot \|_{L_r}) \equiv \log(N(\epsilon, \mathcal{F}_n, \| \cdot \|_{L_r})) \) as a measure of the complexity of \( \mathcal{F}_n \). The second notion of complexity of the class \( \mathcal{F}_n \) is the covering numbers with bracketing. Let \( \mathcal{C}_{L_r} \) be the completion of \( \mathcal{F}_n \) under the norm \( \| \cdot \|_{L_r} \). The \( L_r(P_0) \)-covering numbers with bracketing, denoted by \( N(\epsilon, \mathcal{F}_n, \| \cdot \|_{L_r}) \), is the minimal number of \( N \) for which there exist \( \epsilon \)-brackets \( \left\{ [l_j, u_j] : l_j, u_j \in \mathcal{C}_{L_r}, \max_{1 \leq j \leq N} |l_j - u_j|_{L_r} \leq \epsilon, \| u_j \|_{L_r} < \infty, \| u_j \|_{L_r} < \infty, j = 1, \ldots, N \right\} \) to cover \( \mathcal{F}_n \).

Similarly we let \( H(\epsilon, \mathcal{F}_n, \| \cdot \|_{L_r}) \equiv \log(N(\epsilon, \mathcal{F}_n, \| \cdot \|_{L_r})) \) which is called the \( L_r(P_0) \)-metric entropy with bracketing. See Pollard (1984), Andrews (1994), van der Vaart and Wellner (1996), and van der Geer (2000) for more detailed discussions of metric entropy. We also use a simplified
notation \( b_{1n} \asymp b_{2n} \) for two sequences of positive numbers \( b_{1n} \) and \( b_{2n} \) when the ratio of these two \( b_{1n}/b_{2n} \) is bounded below and above by some positive constants.

Now let \( \mathcal{F}_n = \{ l(y|w, \alpha) - l(y|w, \alpha_0) : \|\alpha - \alpha_0\|_q \leq \delta, \alpha \in \mathcal{A}_n \} \) and for some constant \( b > 0 \), let

\[
\delta_n = \inf \left\{ \delta \in (0, 1) : (\sqrt{n} \delta^2)^{-1} \int_{b_0^2} H(\epsilon, \mathcal{F}_n, \|\cdot\|_{L_2}) d\epsilon \leq \text{const} \right\} .
\] (23)

Note that \( \delta_n \) does not only depend on the smoothness of \( l(\cdot, \cdot) \) but also it depends on the complexity of the sieve \( \mathcal{A}_n \). Now we present Theorem 3.2 in Chen (2007) tailored to our estimator.

**Theorem C.2 (Theorem 3.2 in Chen (2007))**

Suppose that (CC1) The data \( \{Y_{i1}, Y_{i2}, W_i\}_{i=1}^n \) are i.i.d; (CC2) There is a \( C_1 > 0 \) such that for all small \( \epsilon > 0 \), \( \sup_{\alpha \in \mathcal{A}_n, \|\alpha - \alpha_0\|_{L_2} \leq \epsilon} \text{Var} (l(Y_i|W_i, \alpha) - l(Y_i|W_i, \alpha_0)) \leq C_1 \epsilon^2 \); (CC3) For any \( \delta > 0 \), there exists a constant \( \iota \in (0, 2) \) such that \( \sup_{\alpha \in \mathcal{A}_n, \|\alpha - \alpha_0\|_{L_2} \leq \delta} \|l(Y_i|W_i, \alpha) - l(Y_i|W_i, \alpha_0)\| \leq \delta^\iota \|U(W_i)\| \) with \( E \left[ \|U(W_i)\|^{\iota} \right] \leq C_2 \) for some \( t \geq 2 \).

Let \( \tilde{\alpha}_n \) be the approximate sieve ML defined in (15).

Then, \( \|\tilde{\alpha}_n - \alpha_0\|_2 = O_p(\epsilon_n) \) with \( \epsilon_n = \max\{\delta_n, \|\Pi_n \alpha_0 - \alpha_0\|_2\} \).

Note that \( \delta_n \) increases with the complexity of the sieve space \( \mathcal{A}_n \), which can be interpreted as a measure of the standard deviation, while we interpret the approximation error \( \|\Pi_n \alpha_0 - \alpha_0\|_2 \) as a measure of the bias since it decreases with the complexity of \( \mathcal{A}_n \). Now we prove Proposition 4.2 by showing the conditions in Theorem C.2 hold under Assumptions ID, SA2, and SA3. We impose the following conditions

**Condition 2**

(i) \( \sup_{\alpha \in \mathcal{A}_n, \|\alpha - \alpha_0\| = o(1)} |M^{(t)}_{ij}(W, \alpha)| \leq C_{11}(W) < \infty \), \( \forall \alpha \in \{\phi_1, \phi_2, \beta_1, \beta_2, h\}, \forall i, j = 0, 1; \)

(ii) \( \inf_{\alpha \in \mathcal{A}_n, \|\alpha - \alpha_0\| = o(1)} |M^{(t)}_{ij}(W, \alpha)| \geq C_{12}(W) > 0 \), \( \forall \alpha \in \{\phi_1, \phi_2, \beta_1, \beta_2, h\}, \forall i, j = 0, 1; \)

**Condition 3**

\( \sup_{\alpha \in \mathcal{A}_n, \|\alpha - \alpha_0\| = o(1)} |M^{(t)}_{ij}(W, \alpha) - M^{(t)}_{ij}(W, \alpha_0)| = C_2(W) \|\alpha - \alpha_0\|_s \) with \( E[C_2^2(W)] < \infty \), \( \forall \alpha \in \{\phi_1, \phi_2, \beta_1, \beta_2, h\}, \forall i, j = 0, 1. \)

These conditions are straightforward to verify as we discuss in Section E. These conditions are sufficient to verify Conditions (CC2) and (CC3) as we show below.

**C.3.1 Proof of Proposition 4.2**

We prove Proposition 4.2 by verifying conditions in Theorem C.2. Condition (CC1) is directly assumed by Assumption SA3 (i). Condition 1 implies that the pathwise derivative of \( l(\cdot, \cdot) \) is well defined. Conditions 1 and 2 imply that \( l(\cdot, \cdot) \) satisfies a Lipschitz condition on \( \mathcal{A} \). Condition 3 is useful to provide some regularities on the difference of the pathwise derivatives of \( l(\cdot, \cdot) \). We first verify Condition (CC2). Using the mean value theorem, we have \( l(y|w, \alpha) - l(y|w, \alpha_0) = \frac{d l(y|w, \tilde{\alpha})}{d \alpha} [\alpha - \alpha_0] \) where \( \tilde{\alpha} \) lies between \( \alpha \) and \( \alpha_0 \). It follows that \( E \left[ (l(Y|W, \alpha) - l(Y|W, \alpha_0))^2 \right] = \)
\[ E \left[ \left( \frac{dl(W, \alpha)}{d\alpha} \right)^2 \right] \] 

from which we conclude \[ E \left[ \left( (l(W, \alpha) - l(W, \alpha_0))^2 \right) \right] \approx \| \alpha - \alpha_0 \|_2 \] by definition of the weak metric \(| \|_2 \) and by Conditions 2-3. Thus, Condition (CC2) holds immediately.

Now we need to calculate \( l(W, \alpha) - l(W, \alpha_0) \) for all small \( \varepsilon \). This fact and the fact that \( L_{ij}(W, \alpha) \) satisfies the Lipschitz condition for all \( i, j = 0, 1 \) (Conditions 1 and 2) imply that

\[ |l(W, \alpha) - l(W, \alpha_0)| \leq C(W) \| \alpha - \alpha_0 \|_s \tag{24} \]

with \( E \left[ C(W)^2 \right] < \infty \). By Theorem 1 of Gabushin (1967) (for integer \( \nu_1 > 0 \)) or Lemma 2 in Chen and Shen (1998) (for any positive number \( \nu_1 \)), we have

\[ \| \alpha - \alpha_0 \|_s \leq C_1 \| \alpha - \alpha_0 \|_2^{2\nu_1/(2\nu_1 + d_s)}. \tag{25} \]

From (24) and (25), we see that Condition (CC3) is satisfied with \( i = 2\nu_1/(2\nu_1 + 1) \) and \( u(W) = C_1 C(W) \). We have verified all conditions in Theorem C.2.

The next step is to derive the convergence rate depending on the choice of sieves. For the sieve \( A_n \equiv \Theta \times \mathcal{H}_n \) with \( \mathcal{H}_n \) defined in (13), we have \( \| \Pi_n \alpha_0 - \alpha_0 \|_s = O \left( K_n^{-\nu_1/d_s} \right) \) by Lorentz (1986).

Now we need to calculate \( \delta_n \) that solves (23). Now let \( C = \sqrt{E \left[ U(W) \right] } \) and \( u_h = \sup_{h \in \mathcal{H}_n} \| h \|_\infty \), then for all \( 0 < \frac{\varepsilon}{K_n} \leq \delta < 1 \), we have \( H \left( \epsilon, \mathcal{F}_n, \| \cdot \|_{L_2} \right) \leq \log N(\epsilon/C, \mathcal{H}_n, \| \cdot \|_\infty) \leq const \times K_n \times \log (1 + 4u_h/\varepsilon) \) by Lemma 2.5 in van der Geer (2000). Using this result, we obtain

\[ C \frac{1}{\sqrt{2\pi n}} \int_{\delta_n}^{\delta_n^2} \sqrt{H \left( \epsilon, \mathcal{F}_n, \| \cdot \|_{L_2} \right) } d\epsilon \leq \frac{1}{\sqrt{2\pi n}} \int_{\delta_n}^{\delta_n^2} \sqrt{K_n \times \log (1 + 4u_h/\varepsilon)} d\epsilon \leq C_1 \frac{1}{\sqrt{2\pi n}} \sqrt{K_n \delta_n} \leq const. \tag{26} \]

From the last inequality of (26), we obtain \( \delta_n \approx \sqrt{K_n/n} \). We then complete the proof by combining this result with \( \| \Pi_n \alpha_0 - \alpha_0 \|_s = O \left( K_n^{-\nu_1/d_s} \right) \). Finally, by letting \( \delta_n \approx \| \Pi_n \alpha_0 - \alpha_0 \|_s \) and \( K_n = n^{\alpha} \), we obtain the optimal rate with \( K_n = n^{1/(2\alpha_1/d_s + 1)} \).

### C.4 Asymptotic Normality

We derive the asymptotic normality of the structural parameter estimates using Theorem 4.3 in Chen (2007). Below let \( \epsilon_n \) denote any sequence satisfying \( \epsilon_n = o(n^{-1/2}) \) and let \( \mu_n(g(Y, W)) = n^{-1} \sum_{i=1}^{n} \{ g(Y_i, W_i) - E[g(\cdot)] \} \) denote the empirical process indexed by the function \( g \). The following theorem, tailored to our estimator, shows that the plug-in sieve conditional ML estimator \( f(\hat{\alpha}_n) \) achieves the \( \sqrt{n} \)-asymptotic normality.

**Theorem C.3 (Theorem 4.3 in Chen (2007))**

Suppose that (AN1) (i) there is an \( \omega > 0 \) such that

\[ \left| f(\alpha) - f(\alpha_0) - \frac{df(\alpha)}{d\alpha} [\alpha - \alpha_0] \right| = O \left( \| \hat{\alpha}_n - \alpha_0 \|_2^2 \right) \]

uniformly over \( \alpha \in A_n \) with \( \| \alpha - \alpha_0 \|_2 = o(1) \); (ii) \( \left\| \frac{df(\alpha)}{d\alpha} \right\| < \infty \); (iii) there is a \( \Pi_n v^* \in A_n \) such that \( \| \Pi_n v^* - v^* \|_2 \times \| \hat{\alpha}_n - \alpha_0 \|_2 = o_p(n^{-1/2}) \);
(AN2) \( \sup_{\alpha \in A_n, \|\alpha - \alpha_0\|_2 \leq \delta_n} \mu_n \left( l(Y|W, \alpha) - l(Y|W, \alpha \pm \epsilon_n \Pi_n v^*) - \frac{d l(Y|W, \alpha_0)}{d \alpha} [\pm \epsilon_n \Pi_n v^*] \right) = O_p(\epsilon_n^2) \);

(AN3) \( K(\alpha_0, \alpha_n) - K(\alpha_0, \alpha_n \pm \epsilon_n \Pi_n v^*) = \pm \epsilon_n \times (\alpha_n - \alpha_0, \Pi_n v^*) + o(n^{-1}) \);

(AN4) (i) \( \mu_n \left( \frac{d l(Y|W, \alpha_0)}{d \alpha} [\Pi_n v^* - v^*] \right) = o_p(n^{-1/2}) \); (ii) \( E \left[ \frac{d l(Y|W, \alpha_0)}{d \alpha} [\Pi_n v^*] \right] = o(n^{-1/2}) \);

(AN5) \( n^1/2 \mu_n \left( \frac{d l(Y|W, \alpha_0)}{d \alpha} [v^*] \right) \rightarrow N(0, \sigma^2_\nu^*) \) with \( \sigma^2_\nu^* \equiv \text{Var} \left[ \frac{d l(Y|W, \alpha_0)}{d \alpha} [v^*] \right] > 0 \) for i.i.d data holds and \( ||\alpha_n - \alpha_0||_2^2 = o_p(n^{-1/2}) \).

Then, for the sieve ML estimate \( \alpha_n \) given in (15), we have \( \sqrt{n} (f(\alpha_n) - f(\alpha_0)) \rightarrow N(0, \sigma^2_\nu^*) \).

### C.4.1 Proof of Theorem 4.1

We prove Theorem 4.1 by verifying conditions in Theorem C.3. First note that \( \frac{d f(\alpha)}{d \alpha} [\alpha - \alpha_0] = (\theta - \theta_0)' \lambda \) and hence \( f(\alpha) - f(\alpha_0) - \frac{d f(\alpha)}{d \alpha} [\alpha - \alpha_0] = 0 \). Therefore, Condition (AN1) (i) holds with \( \omega = \infty \). Next, note

\[
\sup_{0 \neq \alpha - \alpha_0 \in \mathbf{V}} \frac{|f(\alpha) - f(\alpha_0)|^2}{\|\alpha - \alpha_0\|^2} = \sup_{0 \neq \alpha - \alpha_0 \in \mathbf{V}} \frac{|\lambda'(\theta - \theta_0)|^2}{(\theta - \theta_0)' \mathbf{E} \left[ \frac{d l(Y|W, \alpha_0)}{d \alpha} [\theta - \theta_0] + \frac{d l(Y|W, \alpha_0)}{d \alpha} [h - h_0] \right]^2} \\
= \sup_{0 \neq (\theta - \theta_0, \theta_0) \in \mathbf{V}} \frac{\lambda'(\theta - \theta_0)^2}{(\theta - \theta_0)' \mathbf{E} \left[ \frac{d l(Y|W, \alpha_0)}{d \alpha} [\theta - \theta_0] + \frac{d l(Y|W, \alpha_0)}{d \alpha} [h - h_0] \right]^2} \\
= \lambda' \left( \mathbf{E} \left[ D_{b'}(Y, W)' D_{b'}(Y, W) \right] \right)^{-1} \lambda
\]

which implies \( f(\alpha) = \lambda' \theta \) is bounded if and only if \( E \left[ D_{b'}(Y, W)' D_{b'}(Y, W) \right] \) is finite positive-definite, in which case we have \( v^* \in \mathbf{V} \) such that \( f(\alpha) - f(\alpha_0) \equiv \lambda'(\theta - \theta_0) = (v^*, \alpha - \alpha_0) \) for all \( \alpha \in \mathbb{A} \) by the Riesz representation theorem. This is satisfied with \( v^* \equiv (v^*_{\theta}, v^*_h) \in \mathbf{V} \) where \( v^*_{\theta} = (E \left[ D_{b'}(Y, W)' D_{b'}(Y, W) \right])^{-1} \lambda \) and \( v^*_h = -b' \times v^*_{\theta} \). Thus, Condition (AN1) (ii) is satisfied by Assumption SA4 (ii). Assumption SA4 (iii) implies that we can find \( \Pi_n v^* \in \mathbb{V}_n \equiv \mathbf{\Theta} \times \mathbf{H}_n \) such that \( \|\Pi_n v^* - v^*\|_s = O(n^{-1/4}) \). Combining this with the condition \( ||\alpha_n - \alpha_0||_2 = o_p(n^{-1/2}) \) from Proposition 4.2, we obtain \( ||\Pi_n v^* - v^*||_2 \times ||\alpha_n - \alpha_0||_2 = o_p(n^{-1/2}) \) with \( n_1/d_2 > 1/2 \). This satisfies Condition (AN1) (iii). Next, we verify Condition (AN3) (we verify Condition (AN2) later). Note that we have

\[
E \left[ \frac{d l(Y|W, \alpha_0)}{d \alpha} [\alpha - \alpha_0] \right] = 0
\]

for any \( \alpha - \alpha_0 \) (it does not need to be in \( \mathbf{V} \)) because (i) the directional derivative of \( l(Y|W, \alpha) \) at \( \alpha_0 \) is well defined and (ii) it is unconstrained maximization (see Shen (1997)). This is the zero expectation of score function (as in a parametric ML). We can show (27) as follows. Denote

\[
\mathbf{L}(Y, W, \alpha) = \exp(l(Y|W, \alpha))
\]

\[
E \left[ \frac{d l(Y|W, \alpha_0)}{d \alpha} [\alpha - \alpha_0] \right] = E \left[ \frac{d \mathbf{L}(Y, W, \alpha_0)}{d \alpha} [\alpha - \alpha_0] \right] = \int \sum_y \frac{d \mathbf{L}(Y = y, W, \alpha_0)}{d \alpha} [\alpha - \alpha_0] \mathbf{L}(Y = y, W, \alpha_0) f_W dW
\]

\[
= \lim_{\tau \rightarrow 0} \frac{d}{d \tau} \int \sum_y \mathbf{L}(Y = y, W, \alpha_0 + \tau(\alpha - \alpha_0)) f_W dW = \lim_{\tau \rightarrow 0} \frac{d}{d \tau} \int f_W dW = 0
\]

where the second equality holds by construction since \( \mathbf{L}(Y = y, W, \alpha_0) = \Pr(Y = y|W, \alpha_0) \) for all \( y \in \{(1,1), (1,0), (0,1), (0,0)\} \), the third equality holds by the interchangeability of integral and derivative and by definition of directional derivative, the fourth equality holds since \( \sum_y \mathbf{L}(Y =
From (27), it also follows that
\[ E \left[ \frac{dI(Y|W, \alpha_0)}{d\alpha} [v^*] \right] = 0 \quad \text{and} \quad E \left[ \frac{dI(Y,W,\alpha_0)}{d\alpha} [\Pi_n v^*] \right] = 0. \quad (28) \]

We also need the following results to verify Condition (AN3). For \( \alpha_3 \in \mathcal{A}_n \) such that \( \|\alpha_3 - \alpha_0\|_2 \leq \delta_n \) and for \( \alpha_1, \alpha_2 \in \mathcal{A}_n - \alpha_0 \), note that
\[
\begin{align*}
&\leq E \left[ \left( \frac{dI(Y|W,\alpha_0)}{d\alpha} [\alpha_1] - \frac{dI(Y|W,\alpha_0)}{d\alpha} [\alpha_2] \right) \frac{\delta n}{\|\alpha_3 - \alpha_0\|_2} \right] \\
&\leq C_1 \|\alpha_3 - \alpha_0\|_s \|\alpha_2\|_s \|\alpha_1\|_s \leq C_1 (\|\alpha_3 - \alpha_0\|_2 \|\alpha_2\|_2 \|\alpha_1\|_2)^{2\nu_1/d_z+1}
\end{align*}
\]

where the first inequality uses the triangle inequality, the second inequality holds by Conditions 2-3, and the last result holds by Theorem 1 of Gabushin (1967) (when \( \nu_1/d_z \) is an integer) or by Lemma 2 in Chen and Shen (1998) for any \( \nu_1/d_z > 0 \). Thus, the condition A4 (i) in Wong and Severini (1991) holds with \( \nu_1/d_z > 1 \). Note also that
\[
\begin{align*}
&\leq C_1 (\|\alpha_3 - \alpha_0\|_2 \|\alpha_2\|_2 \|\alpha_1\|_2)^{2\nu_1/d_z+1}
\end{align*}
\]

and hence the condition A4 (ii) in Wong and Severini (1991) holds with \( \nu_1/d_z > 1 \). Now consider
\[
\begin{align*}
K(\alpha_0, \alpha_n) - K(\alpha_0, \alpha_n \pm \epsilon_n \Pi_n v^*)
&= E \left[ I(Y|W, \alpha_n \pm \epsilon_n \Pi_n v^*) - I(Y|W, \alpha_n) \right]
\end{align*}
\]

where the first inequality holds by definition of \( K(\cdot, \cdot) \), the second equality is using the mean value theorem with \( \bar{\epsilon}_n = o(n^{-1/2}) \), the third equality is obtained using the Taylor expansion, the fourth equality is obtained using \( E \left[ \frac{dI(Y|W,\alpha_0)}{d\alpha} [\pm \epsilon_n \Pi_n v^*] \right] = \pm \epsilon_n E \left[ \frac{dI(Y|W,\alpha_0)}{d\alpha} [v^*] \right] = 0 \) by (28) and using (29) with \( \bar{\epsilon}_n, \epsilon_n = o(n^{-1/2}) \), the last result is from the fact that \( \langle \alpha_1, \alpha_2 \rangle = -E \left[ \frac{d^2I(Y|W,\alpha_0)}{d\alpha^2} [\alpha_1, \alpha_2] \right] \) (see Wong and Severini (1991)). Therefore, Condition (AN3) holds since \( \epsilon_n \) is arbitrary. We turn to Condition (AN4). Note that Condition (AN4) (ii) immediately holds since \( E \left[ \frac{dI(Y|W,\alpha_0)}{d\alpha} [\Pi_n v^*] \right] = 0 \) by (28). Define
\[
M^{(h)}(Y, W, \alpha_0) = \left\{ \begin{array}{l}
Y_1 Y_2 \frac{M_{11}^{(h)}(W, \alpha_0)}{L_{11}} + Y_1 (1 - Y_2) \frac{M_{10}^{(h)}(W, \alpha_0)}{L_{10}} \\
+ (1 - Y_1) Y_2 \frac{M_{01}^{(h)}(W, \alpha_0)}{L_{01}} + (1 - Y_1) (1 - Y_2) \frac{M_{00}^{(h)}(W, \alpha_0)}{L_{00}}
\end{array} \right\}.
\]

From this, it follows that
\[
\frac{1}{n} \sum_{i=1}^{n} \frac{dI(Y_i|W_i, \alpha_0)}{d\alpha} [\Pi_n v^* - v^*] = \frac{1}{n} \sum_{i=1}^{n} \frac{dI(Y_i|W_i, \alpha_0)}{d\alpha} [\Pi_n v^*_h - v^*_h]
\]

and
\[
= \frac{1}{n} \sum_{i=1}^{n} M^{(h)}(Y_i, W_i, \alpha_0) (\Pi_n v^*_h(z_i) - v^*_h(z_i))
\]
and thus Condition (AN4) (i) holds by the Chebyshev inequality from $0 = E \left[ \frac{d(Y_jW_{i\alpha})}{d\alpha} \right] \Pi_n v^* - v^*$
due to (28), and from $\|\Pi_n v^*_h - v^*_h\|_\infty = o(n^{-1/4})$ by Assumption SA4 (iii). Now note that
$$
\frac{d(Y|W_{\alpha0})}{d\alpha} [v^*_h] = \frac{\partial(Y|W_{\alpha0})}{\partial \alpha} v^*_h + \frac{d(Y|W_{\alpha0})}{d\alpha} [v^*_h] = \left( \frac{\partial(Y|W_{\alpha0})}{\partial \alpha} - M^{(b)}(Y, W, \alpha_0) b^* \right) \left( E \left[ D_{b^*}(Y, W)' D_{b^*}(Y, W) \right] \right)^{-1} \lambda
= \left( \frac{\partial(Y|W_{\alpha0})}{\partial \alpha} - \frac{d(Y|W_{\alpha0})}{d\alpha} [b^*] \right) \left( E \left[ D_{b^*}(Y, W)' D_{b^*}(Y, W) \right] \right)^{-1} \lambda
= D_{b^*}(Y, W) \left( E \left[ D_{b^*}(Y, W)' D_{b^*}(Y, W) \right] \right)^{-1} \lambda
$$
where the second equality is obtained using (30) and by definitions of $v^*_h$ and $v^*_h$. This implies that Condition (AN5) is satisfied by the Lindeberg and Levy CLT since $E \left[ \frac{d(Y|W_{\alpha0})}{d\alpha} [v^*_h] \right] = 0$ and
$$
E \left[ \left\| \frac{d(Y|W_{\alpha0})}{d\alpha} [v^*_h] \right\|^2 \right] = E \left[ \lambda' \left( E \left[ D_{b^*}(Y, W)' D_{b^*}(Y, W) \right] \right)^{-1} D_{b^*}(Y, W)' D_{b^*}(Y, W) \left( E \left[ D_{b^*}(Y, W)' D_{b^*}(Y, W) \right] \right)^{-1} \lambda \right]
= \lambda' \left( E \left[ D_{b^*}(Y, W)' D_{b^*}(Y, W) \right] \right)^{-1} \lambda < \infty
$$
by Assumption SA4 (ii). Now note $\sigma_n^2 \equiv \text{Var} \left( \frac{d(Y|W_{\alpha0})}{d\alpha} [v^*_h] \right) = \lambda \left( E \left[ D_{b^*}(Y, W)' D_{b^*}(Y, W) \right] \right)^{-1} \lambda'$. Therefore, the conclusion $\sqrt{n} (\hat{\theta}_n - \theta_0) \rightarrow_d N \left( 0, \left( E \left[ D_{b^*}(Y, W)' D_{b^*}(Y, W) \right] \right)^{-1} \right)$ follows since $\lambda$ is arbitrary with $\lambda \neq 0$.

Now Condition (AN2) remains to be shown. Note that Condition (AN2) is implied by
$$
\sup_{\alpha \in A_n, \|\alpha - \alpha_0\| \leq \delta_n} \mu_n \frac{d(Y|W_\alpha)}{d\alpha} \Pi_n v^* - \frac{d(Y|W_{\alpha0})}{d\alpha} \Pi_n v^* = o_p(n^{-1/2}). \quad (31)
$$
Let $\mathcal{F} = \{ \frac{d(y|w, \alpha)}{d\alpha} \Pi_n v^* : \alpha \in A \}$. Condition 3 implies that $\frac{d(y|w, \alpha)}{d\alpha} \Pi_n v^*$ satisfies the Lipschitz condition with respect to $\alpha$ and the metric $\|\cdot\|_s$. Thus, $\frac{d(y|w, \alpha)}{d\alpha} \Pi_n v^*$ satisfies the condition (3.1) of Theorem 3 in Chen, Linton, and van Keilegom (2003). Note also that $\Theta$ is compact and $\mathcal{H}$ is a subset of a Hölder space. Thus, from the Lipschitz condition and the remark 3 (ii) of Chen, Linton, and van Keilegom (2003, it follows that $\int_0^\infty \log N_{\|\cdot\|, \mathcal{F}, \|\cdot\|_{L_2(f_\alpha)}} \, de < \infty$ by the proof of Theorem 3 in Chen, Linton, and van Keilegom (2003). Now note
$$
E \left[ \left\| \frac{d(Y|W_\alpha)}{d\alpha} \Pi_n v^* - \frac{d(Y|W_{\alpha0})}{d\alpha} \Pi_n v^* \right\|^2 \right] \leq C \cdot E \left[ \left\| \frac{d(Y|W_\alpha)}{d\alpha} \Pi_n v^* - \frac{d(Y|W_{\alpha0})}{d\alpha} \Pi_n v^* \right\| \right]
\times \sup_{w \times y \in \mathcal{W} \times \mathcal{Y}, \|\alpha - \alpha_0\| \leq \delta_n, \alpha \in A_n} \| \frac{d(y|w, \alpha)}{d\alpha} \Pi_n v^* - \frac{d(y|w, \alpha_0)}{d\alpha} \Pi_n v^* \| \rightarrow 0
$$
as $\|\alpha - \alpha_0\|_s \rightarrow 0$ where the last result holds by Condition 3. Therefore, applying Lemma 1 in Chen, Linton, and van Keilegom (2003), we find that (31) holds by combining above two results. This completes the proof.

C.4.2 Proof of Proposition 4.3

Similarly to the proof of Theorem 5.1 in Ai and Chen (2003) we can prove Proposition 4.3. Note that
$$
\sum_{i=1}^n \left( \frac{d(y_i|w_i, \alpha_0)}{d\alpha_j} - \frac{d(y_i|w_i, \alpha_0)}{d\alpha} [b_j] \right)^2
$$
is globally convex in $b_j$ and hence the solution of (17) $\hat{b}_j^*$ must be
bounded by $|\hat{b}_j^s|_s \leq C$. Thus, we can focus on the subset $\{b \in \mathcal{B} : \|b\|_s \leq C\}$. Note that uniformly over $b_j \in \mathcal{H}_n$, $|b_j|_s < C$, we have
\[
\frac{1}{n} \sum_{i=1}^{n} \left( D_{b_j} (y_i, w_i, \hat{\alpha}_n) \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left( D_{b_j} (y_i, w_i, \alpha_0) \right)^2 + o_p(1)
\]
since $D_{b_j} (y, w, \alpha) = D_{b_j} (y, w, \alpha_0) + o_p(1)$ uniformly over $\|\alpha - \alpha_0\|_s = o(1)$ by Condition 3 and $\|b_j\|_s < C$. Thus, it suffices to show that $|\hat{b}_j^s(\cdot) - b_j^s(\cdot)|_s = o_p(1)$ because it implies $D_{b_j^s} (y, w, \alpha_0) = D_{b_j} (y, w, \alpha_0) + o_p(1)$. Then combining the above results, we conclude
\[
\frac{1}{n} \sum_{i=1}^{n} \left( D_{b_j} (y_i, w_i, \hat{\alpha}_n) \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left( D_{b_j} (y_i, w_i, \alpha_0) \right)^2 + o_p(1)
\]
from which the claim follows. Finally, note that $|\hat{b}_j^s(\cdot) - b_j^s(\cdot)|_s = o_p(1)$ is satisfied by Condition 3, Assumption SA4, $\|\hat{\alpha}_n - \alpha_0\|_s = o(1)$, and Lemma A.1 in Newey and Powell (2003).

D  Asymptotic Normality for the Type Probability $p$

We can prove Proposition 4.4 by showing all conditions in Theorem 2 of Chen, Linton, and van Keilegom (2003) hold. Here instead we directly prove Proposition 4.4 since it is a simple case of Chen, Linton, and van Keilegom (2003).

D.1  Proof of Proposition 4.4

Let $M(h)$ denote $\int_{\mathcal{Z}} L(h(z)) f_\mathcal{Z}(z) dz$ and $M_n(h)$ denote $\frac{1}{n} \sum_{i=1}^{n} L(h(z_i))$. We have
\[
\sqrt{n} \left( \hat{p}_n - p_0 \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( L(\hat{h}_n(z_i)) - L(h_0(z_i)) \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( L(h_0(z_i)) - E[L(h_0(Z_i))] \right)
\]
\[
= \sqrt{n} \left( M_n(\hat{h}_n) - M_n(h_0) \right) + \sqrt{n} \left( M_n(h_0) - M(h_0) \right). \tag{32}
\]

Now let $\xi(h) = M_n(h) - M(h)$ be a stochastic process indexed by $h \in \mathcal{H}$. Then, we obtain the following stochastic equicontinuity such that for any positive sequence $\delta_n = o(1)$,
\[
\sup_{\|h - h_0\|_\infty \leq \delta_n} |\xi(h) - \xi(h_0)| = o_p(n^{-1/2}) \tag{33}
\]
by applying Lemma 1 of Chen, Linton, and van Keilegom (2003) after establishing that (a) $\{v(h) \equiv L(h) - M(h_0) : h \in \mathcal{H}\}$ is a Donsker class and that (b) $E \left[ (v(h_1) - v(h_2))^2 \right] \to 0$ as $\|h_1 - h_2\|_\infty \to 0$. Consider that $\{v(h) \equiv L(h) - M(h_0) : h \in \mathcal{H}\}$ is a subset of $\Lambda_{\mathcal{C}^4}(\mathcal{Z})$ and $\Lambda_{\mathcal{C}^4}(\mathcal{Z})$ is a Donsker class by Theorem 2.5.6 of van der Vaart and Wellner (1996). Thus, the condition (a) is satisfied. Now note
\[
E \left[ (v(h_1) - v(h_2))^2 \right] = E \left[ (L(h_1) - L(h_2))^2 \right] \leq E \left[ \|L(h_1) - L(h_2)\| \sup_{\|h_1 - h_2\|_\infty} \|L(h_1) - L(h_2)\| \|h_1 - h_2\|_\infty \right]
\]
\[
= E \left[ \|L(h_1) - L(h_2)\| \sup_{\|h_1 - h_2\|_\infty} \|L(h_1) - L(h_2)\| \|h_1 - h_2\|_\infty \right] \leq \frac{1}{4} E \left[ \|L(h_1) - L(h_2)\| \|h_1 - h_2\|_\infty \right]
\]

since $L' = L(1 - L) \leq 1/4$ where the second equality is obtained by applying the mean value theorem and thus, the condition (b) is satisfied. Therefore, (33) holds. Now consider
\[
\sqrt{n} \left( M_n(h_n) - M_n(h_0) \right)
\]
\[
= \sqrt{n} \left( M(\hat{h}_n) - M(\hat{h}_0) \right) + \sqrt{n} \left( M_n(\hat{h}_n) - M(\hat{h}_n) - M_n(h_0) + M(h_0) \right) \tag{34}
\]
\[
= \sqrt{n} \left( M(\hat{h}_n) - M(h_0) \right) + o_p(1)
\]

35
where the last result is obtained by (33) and \(\|\hat{h}_n - h_0\|_\infty = o_p(1)\). Now applying the mean value theorem, we have

\[
\sqrt{n} \left( M(\hat{h}_n) - M(h_0) \right) = \sqrt{n} \int_Z L'(\hat{h}_n(z)) \left( \hat{h}_n(z) - h_0(z) \right) f_Z(z) \, dz
\]

\[
= \sqrt{n} \int_Z L'(h_0(z)) \left( \hat{h}_n(z) - h_0(z) \right) f_Z(z) \, dz
\]

\[
+ \sqrt{n} \int_Z \left( L'(\hat{h}_n(z)) - L'(h_0(z)) \right) \left( \hat{h}_n(z) - h_0(z) \right) f_Z(z) \, dz
\]

where \(\hat{h}_n\) lies between \(\tilde{h}_n\) and \(h_0\). Applying the mean value theorem again, we obtain (noting \(L' = L(1 - L)\))

\[
|L'(\tilde{h}_n) - L'(h_0)| = |L(\tilde{h}_n)(1 - L(\tilde{h}_n)) - L(h_0)(1 - L(h_0))|
\]

\[
= |(1 - L(\tilde{h}_n) - L(h_0))L'(\tilde{h}_n)(\tilde{h}_n - h_0)| \leq (1/4)|\tilde{h}_n - h_0|
\]

where \(\tilde{h}_n\) lies between \(\tilde{h}_n\) and \(h_0\). It follows that \(\sqrt{n} \int_Z (L'(\tilde{h}_n(z)) - L'(h_0(z)))(\hat{h}_n(z) - h_0(z))f_Z(z) \, dz \leq \frac{1}{4}\sqrt{n}|\tilde{h}_n - h_0|\|f_Z\|_\infty |\tilde{h}_n - h_0| = o_p(1)\) by the condition (i). Thus, we find

\[
\sqrt{n} \left( M(\hat{h}_n) - M(h_0) \right) = \sqrt{n} \int_Z L'(h_0(z))(\hat{h}_n(z) - h_0(z))f_Z(z) \, dz + o_p(1).
\]

From (32), (34), and (36), the claim then follows by the condition (ii).

### D.2 Proof of Proposition 4.5

Let \(M_n^*(h) = \frac{1}{n} \sum_{i=1}^n L(h(z_i^n))\) and \(\xi^*(h) = M_n^*(h) - M_n(h)\). Due to Giné and Zinn (1990), we first note that

\[
\sup_{\|h - h_0\|_\infty \leq \delta_n} |\xi^*(h) - \xi^*(h_0)| = o_p(1).
\]

Write \(\sqrt{n} \left( \hat{\mu}_n^* - \mu_n \right) = \sqrt{n} \left( M_n^*(\hat{h}_n^*) - M_n^*(\hat{h}_n) \right) + \sqrt{n} \left( M_n^*(\hat{h}_n^*) - M_n^*(\hat{h}_n) \right) + o_p(1)\). Also note

\[
\sqrt{n} \left( M_n^*(\hat{h}_n^*) - M_n^*(\hat{h}_n) \right)
\]

\[
= \sqrt{n} \left( M_n(\hat{h}_n^*) - M_n(\hat{h}_n) \right) - \sqrt{n} \left( M_n(\hat{h}_n^*) - M_n(\hat{h}_n) \right)
\]

\[
+ \sqrt{n} \left( M_n(\hat{h}_n^*) - M_n(\hat{h}_n) \right) = \sqrt{n} \left( M_n(\hat{h}_n^*) - M_n(\hat{h}_n) \right) + o_p(1)
\]

where the last equality is obtained using (37) and by the conditions (i) and (ii). Now consider

\[
\sqrt{n} \left( M_n(\hat{h}_n^*) - M_n(\hat{h}_n) \right)
\]

\[
= \sqrt{n} \left( M(\hat{h}_n^*) - M(\hat{h}_n) \right) + \sqrt{n} \left( M_n(\hat{h}_n^*) - M(\hat{h}_n^*) \right) - \sqrt{n} \left( M_n(\hat{h}_n) - M(\hat{h}_n) \right)
\]

\[
+ \sqrt{n} \left( M_n(\hat{h}_n) - M(\hat{h}_n) \right) = \sqrt{n} \left( M(\hat{h}_n^*) - M(\hat{h}_n) \right) + o_p(1)
\]

by the conditions (i), (ii), and (iii). Now similarly with (36), we can show that \(\sqrt{n} \left( M(\hat{h}_n^*) - M(\hat{h}_n) \right) = \sqrt{n} \int_Z L'(h_0)(\hat{h}_n^* - h_0) \, dF_Z + o_p(1)\). From above results the claim then follows by the condition (iv).
D.2.1 Stochastic Expansion of $\hat{h}_n - h_0$

To check the condition (ii) in Proposition 4.4 (or the condition (iv) in Proposition 4.5), we need to derive the stochastic expansion of $\hat{h}_n - h_0$. Here we provide an expansion for the sieve conditional ML estimator. To facilitate this task, we define a pseudo true value of $\theta$ and $h$ such that

$$\alpha_{0K} \equiv (\theta_{0K}, h_{0K}) = \arg\max_{\theta \in \Theta, h = R^K(\cdot) \pi \in \mathcal{H}_n} Q_K(\theta, h) \equiv E [l(Y|W, \theta, h(Z))]$$

(38)

and let $h_{0K}(\cdot) = R^K(\cdot)' \pi_{0K}$. Similarly, we let $\hat{h}_n(\cdot) = R^{K_n}(\cdot)' \pi_n$. Then,

$$\hat{h}_n(\cdot) - h_{0K}(\cdot) = R^K(\cdot)' (\pi_n - \pi_{0K})$$

with $K = K_n$.

Define $\Psi(h) = \int_Z L(h(z))(1 - L(h(z))) R^K(z) dF_Z(z)$.

Suppose (a) $\Psi(h_0)'(\pi_n - \pi_{0K}) = \frac{1}{n} \sum_{i=1}^n \psi(\alpha_{0K}, Y_i, W_i) + o_p(n^{-1/2})$, where $E[\psi(\alpha_{0K}, Y_i, W_i)] = o(1)$ and $E[\|\psi(\alpha_{0K}, Y_i, W_i)\|^2_E] < \infty$ and suppose (b) $\|h_{0K} - h_0\|_\infty = o(n^{-1/2})$. Then, we have

$$\sqrt{n} \int_Z L(h_0)(1 - L(h_0)) \left(\hat{h}_n - h_0\right) dF_Z$$

$$= \sqrt{n} \int_Z L(h_0)(1 - L(h_0)) \left(\hat{h}_n - h_{0K}\right) dF_Z + \sqrt{n} \int_Z L(h_0)(1 - L(h_0)) (h_{0K} - h_0) dF_Z$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(\alpha_{0K}, Y_i, W_i) + o_p(1)$$

where the second equality is obtained from (39) and the definition of $\Psi(h)$, and the third equality is obtained from (a) and (b) noting $L(\cdot)(1 - L(\cdot)) \leq 1/4$. From this, it follows that

$$V_p = \lim_{K \to \infty} E \left[ (\psi(\alpha_{0K}, Y_i, W_i) + \varphi(Z_i)) (\psi(\alpha_{0K}, Y_i, W_i) + \varphi(Z_i))' \right]$$

where $\varphi(Z_i) = L(h_0(Z_i)) - E[L(h_0(Z_i))]$. Therefore, to verify the condition (ii) of Proposition 4.4 holds, it suffices to show (a) and (b) hold. The condition (a) can be verified using the first order conditions of (14) and (38) because with fixed $K$, the problem becomes parametric. For (b), we can show $\|h_{0K} - h_0\|_\infty = \zeta(K) K^{-1/2} \sup_{z \in \mathbb{Z}} \|R^K(z)\|_E$ similarly with Hirano, Imbens, and Ridder (2003) who consider a logit series estimation.

E Smoothness of Conditional Probabilities

For the conditional probabilities presented in Appendix B, we show that the pathwise first and second derivatives are well-defined. This result is useful to verify Conditions 1 and 2-3 for the sieve conditional ML. It is easy to see that the pathwise derivatives are well-defined as long as $G_1(\cdot)$ and $G_2(\cdot)$ are continuously differentiable since the function $h(\cdot)$ appears only in $p(Z) = \exp(h(z))/(1 + \exp(h(z)))$ and we have $\frac{dp(Z)}{dh}[h_1 - h_2] = (1 - p(z))p(z)(h_1 - h_2)$. Therefore, for the conditional probabilities given in Appendix B, we have

$$\frac{dP_{ij}(Y|W, \theta, p(z))}{dh}[h_1 - h_2] = M_{ij}^{(h)}(h_1 - h_2), \forall i, j = 0, 1,$$

$$\frac{d^2P_{ij}(Y|W, \theta, p(z))}{dh^2}[h_1 - h_0, h_2 - h_0] = M_{ij}^{(h)(h)}(h_1 - h_0)(h_2 - h_0), \forall i, j = 0, 1, \text{ and}$$

$$\frac{d\theta P_{ij}(Y|W, \theta, p(z))}{dh\theta}[h_1 - h_2] = M_{ij}^{(h)(t)}(h_1 - h_0), \forall i, j = 0, 1 \text{ and for any element } t \text{ of } \theta,$$
where $M_{ij}^{(h)}$, $M_{ij}^{(h)(h)}$, and $M_{ij}^{(h)(t)}$ are some well-defined derivatives. We also note that those derivatives and other derivatives with respect to finite dimensional parameters are uniformly bounded by some constants when (i) $G_1()$ and $G_2()$ are continuously differentiable, (ii) the parameter space $\Theta$ is compact, (iii) $\mathcal{W}$ is compact, and (iv) $0 < p(Z) < 1$. Therefore, the Lipschitz conditions for the conditional probabilities and the Lipschitz conditions for the pathwise first derivatives of the conditional probabilities are satisfied. This implies that the Lipschitz conditions for the log likelihood and the Lipschitz conditions for the pathwise derivatives of the log likelihood are also well-defined.

References


