Simulated MLE for Discrete Choices using Transformed Simulated Frequencies
Donghoon Lee\(^2\) and Kyungchul Song\(^3\)

May 26, 2011

Abstract

Many methods of simulated likelihood for discrete choice models that have been developed in the literature require the presence of additive errors that have normal or extreme value distributions, with the prominent exception of the original simulated frequency method of Lerman and Manski (1981). This paper proposes a new method of simulated likelihood that is free from simulation bias for each finite number of simulation, and yet flexible enough to accommodate a variety of model specifications beyond those of additive normal or logit errors. The flexibility in modeling comes from the fact that the method depends only on simulated frequencies of individual choices, and hence entails a price: the sample objective function is discontinuous in the parameters. The method begins with the likelihood function involving simulated frequencies and finds a transform of the likelihood function that identifies the true parameter for each finite simulation number. The transform is explicit, containing no unknowns that demand an additional step of estimation. The estimator achieves the efficiency of MLE as the simulation number increases fast enough. This paper presents and discusses results from Monte Carlo simulation studies of the new method.

Key words: Simulated MLE, Discrete Choice Models, Simulation Bias, Simulated Frequencies, Cube-Root Asymptotics

JEL Classifications: C12, C14, C52.

\(^1\) A previous version was titled, "A Consistent SMLE When the Simulation Number was Finite." We are indebted to Sean D. Campbell and Eugene Savin for his valuable comments and inputs. We are also grateful to the seminar participants at Columbia University, Econometric Society Meeting, Greater New York Econometrics Colloquium, Lehigh University, University of Iowa, University of Rochester, and New York University for useful comments. We thank a Co-Editor and two anonymous referees for their thoughtful comments. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of New York or the Federal Reserve System.

\(^2\) Federal Reserve Bank of New York, 33 Liberty Street, New York, NY, 10045.

1 Introduction

Discrete choice models have long been used in a wide range of empirical fields of economics. While a discrete choice model typically specifies the data generating process up to a parametric family of distributions, maximum likelihood estimation is infeasible in practice except for simple models because the explicit evaluation of the likelihood is not possible. Since the seminal work of Lerman and Manski (1981), the approach of simulation-based inference has been increasingly instrumental for overcoming this difficulty, providing the researcher with a wider spectrum of flexibility in modeling. (See Hajivassiliou and Ruud (1994), Stern (1997), Gouriéroux and Monfort (1997), and Train (2003) for a review of the literature and references therein.)

Most developments of methods of simulated likelihood have been made with a requirement that the original method of Lerman and Manski (1981) was free from: the assumption of additive normal or logit errors in the latent processes. For example, while the method of Stern (1992) and the method of GHK simulator (Geweke (1989), Hajivassiliou (1990) and Keane (1993)) are computationally efficient, these methods rely on the assumption that the random utilities involve additive normal errors. Hajivassiliou (1990) and McFadden and Hajivassiliou (1998) proposed a different method of simulated likelihood that uses simulated scores to construct simulated moment conditions and proved efficiency of the estimators. In particular Hajivassiliou and McFadden (1998) suggested three methods of maximum simulated scores (named MSS-SAR, MSS-SRC, and MSS-GRS in their paper). These methods also require that the random utilities have additive multivariate normal errors. Another increasingly popular class of discrete choice models include mixed multinomial logit models (MMLM) (Ben-Akiva, et. al. (1997) and McFadden and Train (2000) and see references therein). The MMLMs offer a flexible way of modeling heterogeneity through random coefficient specifications and yet requires the presence of additive logit errors.

Many structural econometric models used in labor economics and industrial organizations do not admit such simple modeling of random utilities. In these models, unobserved heterogeneity in individual decision making often lies at the center of econometric modeling. Depending on the way one models the role of heterogeneity, one can easily encounter a structural model for which aforementioned simulation methods do not apply. (See Keane and Wolpin (1994), Keane and Wolpin (1997). etc.) In these situations, the original simulated frequency method of Lerman and Manski or their smoothed variants tend to emerge as the sole feasible solution. As is well-documented, however, the simulated frequency method of Lerman and Manski poses several difficulties such as discontinuity of the sample objective function, the zero probability problem, and the simulation bias due to the use of only a finite
number of simulations.

This paper proposes a new method of simulated likelihood (MSL from here on) for discrete choices. Our method does not require additive normal or logit specification of random utilities and flexibly applies to all the models that the procedure of Lerman and Manski applies to. At the same time, the method is free from the zero-probability problem and does not accompany simulation bias for each finite simulation number. (See Lee (1995) for the asymptotic bias analysis of simulated discrete choice models and for a bias adjustment method.) The method is easy to implement, accompanying almost no additional computational cost beyond that of the simulated frequency method of Lerman and Manski. To the best of our knowledge, our method is the first simulated likelihood method that does not suffer from simulation bias for finite simulation numbers and yet allows for flexible modeling beyond that of normal or logit additive errors.

Our method is built on the main finding of this paper that there exists a simple and explicit transform of a simulated likelihood function whose maximization delivers a consistent estimator even with a finite simulation number. The transform is algebraically explicit, depending on no unknowns. Furthermore, the use of the transform does not require any restrictions on the specification of the random utilities, and hence flexibly applies to many discrete choice models that have a nonlinear, nonnormal form of heterogeneity. We call this new method *transformed simulated frequencies* (TSF) method. Our approach, however, shares one drawback of other simulation methods that use simulated frequencies, such as MSL of Lerman and Manski (1981) or methods of simulated moments of McFadden (1989): the sample objective function is discontinuous in parameters. The comparison of our method with other existing ones is summarized in Table 1.

### Table 1: Comparison of Simulated MLE Methods

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Method (TSF)</td>
<td>No</td>
<td>Discontinuous</td>
<td>Not Required</td>
<td>No Bias</td>
</tr>
<tr>
<td>Lerman-Manski</td>
<td>Yes</td>
<td>Discontinuous</td>
<td>Not Required</td>
<td>Bias</td>
</tr>
<tr>
<td>Stern (1992)</td>
<td>No</td>
<td>Smooth</td>
<td>Required (Normal)</td>
<td>Bias</td>
</tr>
<tr>
<td>GHK</td>
<td>No</td>
<td>Smooth</td>
<td>Required (Normal)</td>
<td>Bias</td>
</tr>
<tr>
<td>MSS - SRC/GRS</td>
<td>No</td>
<td>Smooth</td>
<td>Required (Normal)</td>
<td>Bias</td>
</tr>
<tr>
<td>MSS - SAR</td>
<td>No</td>
<td>Discontinuous</td>
<td>Required (Normal)</td>
<td>No Bias</td>
</tr>
<tr>
<td>Mixed Multinomial Logit</td>
<td>No</td>
<td>Smooth</td>
<td>Required (Logit)</td>
<td>Bias</td>
</tr>
</tbody>
</table>

In this paper, we formally present conditions for identification and derive the asymptotic theory for the estimator in both the cases of simulation numbers fixed and increasing with
the sample size. Our exposition is made through easily verifiable, high-level conditions to emphasize the flexibility of our approach. The conditions require only weak regularity conditions for the stochastic link between the decision variables and the observed covariates. We also demonstrate how our framework can also be adapted to the case where only the cohort-level aggregate data are available under certain conditions. This set-up is relevant to some empirical researches in industrial organizations.

Here is the summary of the asymptotic properties of the estimators based on the TSF method. When the simulation number is fixed and the sample size \( n \) increases, the estimator is consistent at the rate of \( \sqrt{n} \), like the maximum score estimator (Manski (1975) and Kim and Pollard (1990)). In the case of an increasing number of simulations, we establish that the estimator is \( \sqrt{n} \)-consistent and asymptotically normal as the simulation number increases to infinity at a rate slightly faster than \( \sqrt{n} \). This latter condition is only slightly stronger than the existing condition for many MSL estimators. (See e.g. Lerman and Manski (1981) and Gouriéroux and Monfort (1997).) Under this same condition, the estimator achieves the asymptotic efficiency of MLE.

To illustrate the usefulness of our approach, we performed Monte Carlo simulation studies based on two types of schooling choice models which involves heterogeneity in discount factor and ability. More specifically, the discount factor is assumed to be correlated with other observed individual characteristics and also an unobserved characteristic. In the second type of models, we assume time-varying heterogeneity so that the econometric model is a dynamic discrete choice model. The simulation methods considered in this study are, Lerman-Manski’s MSL and its smoothed version, because these are the methods that are applicable in these models. Our estimator mostly dominates Lerman and Manski’s simulation method and smoothed versions regardless of the simulation number. The domination is prominent especially when the simulation number is small and the sample size is large.

The remainder of this paper is organized as follows. In section 2, we define the class of discrete choice models, discuss MSL, and offer a preview of our method. In section 3, we present the main results of this paper which formally establish identification and consistency of the proposed estimator. It is also shown that the estimator is asymptotically normal when the simulation number goes to infinity fast enough. In Section 4, we present and discuss results from Monte Carlo simulation studies. Section 5 concludes. All the technical proofs are relegated to the appendix.
2 Discrete Choice Models and TSF

2.1 Methods of Simulated Likelihoods

Suppose that a binary decision variable, \( D_{ij} \in \{0, 1\} \), of an agent \( i \) choosing the \( j \)-th choice, is stochastically linked with an observed covariate vector \( X_i \) as follows:

\[
D_{ij} = \delta_j(X_i, \eta_i; \theta_0),
\]

where \( X_i = (X_{i1}, \cdots, X_{iJ})^\top \) represents a vector of observed covariates, \( \eta_i = (\eta_{i1}, \cdots, \eta_{iJ})^\top \) a vector of unobserved variables, and \( \theta_0 \in \Theta \subset \mathbb{R}^d \) the parameter to be estimated. The number \( J \) denotes the number of the choices the agent encounters and \( n \) the number of the agents in the data set. For example, \( \delta_j \) can be specified as follows,

\[
\delta_j(X_i, \eta_i; \theta_0) = 1 \left\{ u_j(X_i, \eta_i; \theta_0) \geq \max_{1 \leq k \leq J} u_k(X_i, \eta_i; \theta_0) \right\},
\]

where the function \( u_j(X_i, \eta_i; \theta) \) is a random utility (McFadden (1974)).

The conditional choice probability of the \( i \)-th agent choosing the \( j \)-th option is defined by

\[
p_j(X_i, \theta_0) = P \{ D_{ij} = 1 | X_i \}.
\]

The choice probability is obtained by "integrating out" the unobserved variable \( \eta_i \) conditional on the observed covariate \( X_i \). It is interpreted as the probability of the \( j \)-th choice being made by an agent \( i \) with a covariate \( X_i \). Given the choice probabilities \( p_j(X_i, \theta) \), it is natural to form the log-likelihood of a random sample \( \{D_i, X_i\} \) as follows:

\[
l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \log(p_j(X_i, \theta)).
\]

So far as \( p_j(x, \theta) \) can be evaluated, maximum likelihood estimation is straightforward. (e.g. Amemiya (1985).) However, the choice probability is often hard to evaluate, in particular when the number of choices is large and one wants to admit flexibility in specifying the joint distribution of \( \eta_i \).

Methods of simulated likelihood substitute a simulated choice probability for the choice probability to construct a simulated log-likelihood,

\[
l_{n,R}^*(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \log(p_{jR}^*(X_i, \theta)).
\]
The number $R$ represents the repetition number of simulated stochastic variables. A simulated maximum likelihood estimator is defined as a maximizer of the simulated log-likelihood function,

$$
\hat{\theta}_{n,R}^* = \arg \max_{\theta \in \Theta} I_{n,R}^*(\theta).
$$

When $R$ increases with the sample size fast enough, it is known that for most choice probability simulators in the literature, the resulting estimator is consistent.

The original method of Lerman and Manski (1981) uses simulated frequency in constructing $p_{jR}^*(X_i, \theta)$. More specifically, suppose that $R$ number of stochastic errors $\eta_{i,r}^*, r = 1, \ldots, R$, are drawn from the known distribution $F$ of $\eta_i$. We let $\delta_j(X_i, \eta_{i,r}^*; \theta), r = 1, \ldots, R$, (taking values of 0 or 1) denote simulated choices for each value of $\theta$. The simulated frequency of each choice $j$ at simulation number $R$ is defined to be

$$
m_{jR}(X_i, \eta_i^*; \theta) = \sum_{r=1}^{R} \delta_j(X_i, \eta_{i,r}^*; \theta)
$$

where $\eta_i^* = (\eta_{i,1}^*, \ldots, \eta_{i,R}^*)^T$ is a random sample from the distribution $F$ of $\eta_i$. The number $m_{jR}(X_i, \eta_i^*; \theta)$ represents the number of incidences that the $j$-th choice is made by an agent $i$ who has covariates $X_i$ and simulated stochastic errors $\eta_i^*$. From now on, we write briefly

$$
m_{ij}(\theta) = m_{jR}(X_i, \eta_i^*; \theta),
$$

and $m_i(\theta) = (m_{i1}(\theta), \ldots, m_{iJ_i}(\theta))^T$. Then the simulated choice probability is defined to be

$$
p_{jR}^*(X_i, \theta) = \frac{m_{ij}(\theta)}{R}.
$$

Plugging this into (3), one obtains the MSL estimator of Lerman and Manski.

This simulated frequency method of Lerman and Manski has been known to suffer from several drawbacks. Among which are the zero probability problem, discontinuity of the sample objective function in the parameters and simulation bias. First, the zero probability problem refers to the fact that for small $R$ and large $n$, some of $p_{jR}^*(X_i, \theta)$ is likely to assume a zero value, causing a log-zero problem in the optimization. Second, the procedure involves a sample objective function that is discontinuous in the parameters. Finally, simulation bias occurs because the choice probability $p_{jR}^*(X_i, \theta)$ is different from $p_{jR}^*(X_i, \theta)$.

Many developments since Lerman and Manski (1981) have focused on overcoming the first two problems, i.e., zero probability problem and the problem of discontinuous objective functions (e.g. GHK method, MSS, and MMNL modeling mentioned in the introduction).
Unlike the original method of Lerman and Manski (1981), these developments usually assume that the random utilities involve additive normal or logit errors. In many structural models of labor economics and industrial organizations, however, such specifications are not permitted due to the structural nature of the model. (See e.g. Keane and Wolpin (1994). See also the Monte Carlo simulation examples of this paper.) This paper’s proposal is an alternative that retains the modeling flexibility of the original Lerman-Manski’s procedure, and yet is free from the zero probability problem and simulation bias.

2.2 Transformed Simulated Frequency (TSF)-Based Method

The TSF method begins by attempting to overcome the problem of simulation bias. Certainly, the fact that the simulated choice probabilities are unbiased estimators of the true choice probabilities does not help due to the presence of logarithm in (3). This paper proposes an alternative function different from logarithm that eliminates the simulation bias entirely for each finite simulation number. More specifically, for each fixed $R$, this paper develops a transform $T_{R,j}(\cdot)$ of simulated frequencies, $j = 1, 2, \ldots, J, R = 2, \ldots$, such that

$$\theta_0 = \arg \max_{\theta \in \Theta} \sum_{j=1}^{J} E[D_{ij}T_{R,j}(m_i(\theta))],$$  \hfill (5)

i.e., the population objective function identifies $\theta_0$ for each $R = 2, \ldots$. Then for each $R$, an estimator of $\theta_0$ is naturally obtained by maximizing its sample analogue:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij}T_{R,j}(m_i(\theta)).$$  \hfill (6)

A transform that satisfies (5) under regular conditions turns out to be of the following form: for each $j = 1, 2, \ldots, R$, and $R = 2, 3, \ldots$,

$$T_{R,j}(m) = - \sum_{s=0}^{R-m_j-1} \frac{1}{R-s} + \frac{1}{R} \sum_{k=1,k \neq j}^{J} 1\{m_k > 0\}, \ m = (m_1, \ldots, m_J)^\top,$$  \hfill (7)

where $m_j$’s are nonnegative integers such that $\sum_{j=1}^{J} m_j = R$. The remarkable aspect of the transform $T_{R,j}(\cdot)$ is that the transform does not depend on any unknown aspects of the data generating process. The transform depends only on $J$ and $R$ which are known. This means that we do not have to estimate the transform $T_{R,j}$ when one solves (6). We call $T_{R,j}(m_i(\theta))$ a transformed simulated frequency (TSF).

The main advantage of the TSF-based method comes from the fact that it relies only
on the elementary method of simulated frequencies, and hence does not require a particular
structure of random utilities. Furthermore, the TSF-based is free from the zero probabil-
ity problem by design. In other words, when the sample size is large, using only a finite
simulation number does not cause a zero probability problem, in contrast to the simulated
frequency method of Lerman-Manski. Some simulation results to be presented in the next
subsection illustrate this point.

### 2.3 Illustration

To illustrate how the transform $T_{R,j}$ works, let us consider the following simple simulation
example. We consider a binary choice model where the conditional choice probability of the
first choice given $X \in \mathbf{R}$ is specified as

$$p(X; \theta_0) = \frac{1}{1 + \exp((10 + \theta_0)X)},$$

where $X$ is drawn from the uniform distribution on $[-1, 1]$ and the true parameter $\theta_0$ is set
to be zero.

Figure 1 shows three population objective functions against different values of $\theta$ with
different simulation numbers $R$:

$$Q_{TSF}^R(\theta) = \sum_{j=1}^{J} \mathbf{E}[D_{ij}T_{R,j}(m_i(\theta))]: \text{(TSF)}$$

$$Q_{L-M}^R(\theta) = \sum_{j=1}^{J} \mathbf{E}\left[D_{ij}\log\left(\frac{m_{ij}(\theta)}{R}\right)\right]: \text{(Lerman-Manski)}$$

$$Q_{MLE}(\theta) = \sum_{j=1}^{J} \mathbf{E}[D_{ij}\log(p_j(X_i; \theta))]: \text{(MLE)}.$$

Each panel plots the three population objective functions over different simulation num-
bers $R$. Even when $R = 2$, the maximizers of $Q_{TSF}^R(\theta)$ and $Q_{MLE}(\theta)$ coincide, and this
coincidence is maintained as $R$ increases. When $R$ is large, both the objective functions
$Q_{TSF}^R(\theta)$ and $Q_{MLE}(\theta)$ coincide for all the values of $\theta$. This makes contrast with the objective
function of Lerman-Manski. When $R$ is small, the objective function of Lerman-Manski
is away from the true value $\theta_0 = 0$. This reflects the well-known fact that the MSL estimator
of Lerman-Manski is inconsistent for a finite $R$. Only when $R$ becomes large, Lerman-Manski
objective function becomes close to the true MLE objective function.

Unlike Lerman-Manski, the approach of TSF does not suffer from the zero-probability
Figure 1: Population Objective Functions: The objective function of TSF-based MSL has the same maximizer as that of MLE for each simulation number.
Figure 2: Value of Log of Simulated Probabilities (Lerman and Manski): The plots are against the given true choice probability $p$. The expected value of TSF is bounded from below when $p$ is close to zero, whereas the expected value of log of simulated probabilities falls to $-\infty$.

problem. With each finite sample size $n$ and finite simulation number $R$, TSF $T_{R,j}(m_j(\theta))$ always assumes a finite number regardless of the realizations of the simulated frequency $m_j(\theta)$. Hence the finite sample objective function is well defined regardless of the sample size and the simulation number. To illuminate this point, Figure 2 plots $\log(p)$, $E \log(m_i(\theta_0)/R)$, and $E T_{R,j}(m_i(\theta_0))$ against $p$, the choice probability, where the expected value is over the distribution of simulated errors when $R$ is finite. Here the simulated frequencies $m_i(\theta_0)$ are generated according to the given value of the choice probability $p$. Certainly, in the case of Lerman-Manski, the zero-probability problem is severe when the simulation number is small, as shown by steeply falling curves as we move $p$ close to zero. In contrast, the expected TSF does not suffer from this zero-probability problem. Furthermore, the expected TSF becomes close to $\log(p)$ more quickly than the expected logarithm of simulated probabilities as the simulation number increases.
3 Main Results

3.1 Identification

In this section, we provide the main result of this paper that the use of TSF in (7) identifies \( \theta_0 \) for each finite simulation number \( R \). Let \( m_{ij}(\theta) \) be as defined in (4) and \( \hat{\theta} \) be as defined in (6). We introduce the following regularity conditions.

**Assumption 1:**

(i) \( \Theta \) is compact with a nonempty interior containing \( \theta_0 \) and for all \( \theta \in \Theta \) and \( x \) in the support of \( X_i \), the choice probability \( p_j(x; \theta) \) belongs to \( S_J \in \{ p \in [0, 1]^J : \Sigma_{j=1}^{J} p = 1 \} \).

(ii) For all \( \theta \in \Theta \) such that \( \theta \neq \theta_0 \), there exists \( j \in \{1, \cdots, J\} \) such that \( P \{ p_j(X_i; \theta) \neq p_j(X_i; \theta_0) \} > 0 \).

(iii) For some \( \delta > 0 \), \( \inf_{\theta \in B(\theta_0, \delta)} \inf_{x \in X} p_j(x; \theta_0) > \varepsilon_p > 0 \), \( j = 1, \cdots, J \), for some \( \varepsilon_p > 0 \).

Conditions in Assumption 1 are standard in the MLE of discrete choice models. Condition (i) requires that the choice probability function \( p(x; \theta) \) is well-defined for all \( \theta \in \Theta \). Condition (ii) is a condition used to identify \( \theta_0 \) from the identification of choice probabilities. Condition (iii) requires that the choice probabilities be bounded away from zero, and implies that \( p_j(x; \theta_0) < 1 - \varepsilon_p \) for each \( j = 1, \cdots, J \).

The following theorem is the main theorem of this paper.

**Theorem 1 (Identification):** Suppose that Assumption 1 holds. Then for each \( \delta > 0 \), and for each \( R = 2, 3, \cdots \),

\[
\sum_{j=1}^{J} E[D_{ij}T_{R,j}(m_i(\theta_0))] > \max_{\theta \in \Theta \setminus B(\theta_0, \delta)} \sum_{j=1}^{J} E[D_{ij}T_{R,j}(m_i(\theta_0))],
\]

where \( B(\theta_0, \delta) = \{ \theta \in \Theta : ||\theta - \theta_0|| < \delta \} \).

The identification result in Theorem 1 tells us that in contrast to the use of logarithm, the use of \( T_{R,j} \) leads to identification of \( \theta_0 \) for each finite \( R = 2, 3, \cdots \). This fact forms the basis on which we develop an MSL estimator that is consistent even if we have a finite simulation number \( R \).
3.2 A Heuristic Exposition on the Discovery of TSF

In this section, we explain the way the transform (7) is discovered. Let \( N_R = \{0, 1, 2, \cdots, R\} \) and define

\[
N_{R,J} = \{ (m_1, \cdots, m_J) : m_j \in N_R, \ j = 1, \cdots, J, \ \text{and} \ \sum_{j=1}^J m_j = R \}.
\]

The set \( N_{R,J} \) denotes the space of \( J \)-tuples where simulated frequencies \( m_i(\theta) \) realize. Also we write the conditional choice probability \( p_i(\theta) = p(X_i; \theta) \) for brevity.

To find the right map \( T_{R,J} \), we first focus on some necessary conditions that such a map should satisfy. Given a generic map \( T_j(\cdot) : N_{R,J} \rightarrow \mathbb{R}_+ \), \( j = 1, \cdots, J \), and \( R = 2, 3, \cdots \), we introduce a function \( \Lambda_R(p, p_0; T) : S_J \times S_J \rightarrow \mathbb{R}, T = (T_1, \cdots, T_J) \), as follows:

\[
\Lambda_R(p, p_0; T) = \sum_{j=1}^J p_{j0} \int T_j(M) dF_{R,p}(M),
\]

where \( F_{R,p} \) is the CDF of the multinomial distribution on \( N_{R,J} \) with parameter \((R, p)\). The function \( \Lambda_R \) is uniquely determined once \( R \) and the transform \( T \) are chosen, and it does not depend on any other specifics of the data generating process of \((D_i, X_i)\).

Using \( \Lambda_R \), we rewrite

\[
\sum_{j=1}^J \mathbb{E} [D_{ij} T_j(m_i(\theta))] = \mathbb{E} [\Lambda_R(p_i(\theta), p_i(\theta_0); T)].
\]

The main idea of this paper is that we extract conditions for \( \Lambda_R \) such that

\[
\theta_0 = \arg \max_{\theta \in \Theta} \mathbb{E} [\Lambda_R(p_i(\theta), p_i(\theta_0); T)]. \tag{8}
\]

Invoking the interchangeability of the derivative and the expectation, we write the first order condition for (8) as

\[
\frac{\partial}{\partial \theta} \mathbb{E} [\Lambda_R(p_i(\theta), p_i(\theta_0); T)] \bigg|_{\theta = \theta_0} = \sum_{j=1}^J \mathbb{E} \left[ \lambda_j(p_i(\theta_0), p_i(\theta_0); T) \frac{\partial p_j(X_i, \theta_0)}{\partial \theta} \right] = 0, \tag{9}
\]

where for each \( j = 1, 2, \cdots, J \),

\[
\lambda_j(p, p_0; T) = \frac{\partial \Lambda_R(p, p_0; T)}{\partial p_j} \tag{10}.
\]

Since the choice probabilities \( p_j(x; \theta) \) sum up to one for all \( x \) and \( \theta \), differentiability of \( p_j(x, \theta) \)
at each $\theta$ implies
\[
\sum_{j=1}^{J} \frac{\partial p_j(x, \theta)}{\partial \theta} = 0. \tag{11}
\]
This means that the first order condition in (9) immediately follows if $\lambda_j(p_0, p_0; T)$ is the same across $j$’s, i.e., for each $p_0 \in S_J$,
\[
\lambda_j(p_0, p_0; T) = \lambda_k(p_0, p_0; T), \quad \text{for all } j, k = 1, 2, \cdots, J. \tag{12}
\]
The condition in (12) has an important merit of not depending on any aspects of the data generating process other than what is already fully known, i.e., $R$ and $J$. It remains to search for $T$ such that (12) is satisfied.

We first write out
\[
\Lambda_R(p, p_0; T) = \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,J}} T_j(m) p_R(m; p), \tag{13}
\]
where $p_R(\cdot; p)$ is the probability mass function of $F_{R,p}$ which is given by
\[
p_R(m; p) = \begin{pmatrix} R \\ m_1, \cdots, m_J \end{pmatrix} p_1^{m_1} \cdots p_J^{m_J}.
\]
From (13), $\Lambda_R(p, p_0; T)$ is a weighted sum of $p_R(m; p)$ over $m \in \mathbb{N}_{R,J}$. Hence for each $k$, we can write $\lambda_k(p_0, p_0; T)$ also as a weighted sum of $p_R(m, p_0)$ by rearranging the terms.

For simplicity, let us consider the case of $J = 2$. Since for all $m = (m_1, m_2) \in \mathbb{N}_{R,J}$, $m_1 + m_2 = R$, it suffices to find a map $T$ on $\mathbb{N}_R$. The conditions in (12) imply that the coefficient of $p_R(m_1, m_2; p_0)$ in the sum defining $\lambda_1(p_0, p_0; T)$ is equal to the coefficient of $p_R(m_1, m_2; p_0)$ in the sum defining $\lambda_2(p_0, p_0; T)$ for all nonnegative integers $m_1, m_2$ such that $m_1 + m_2 = R$. From this, one can derive the following difference equations:
\[
m_1 \{T(m_1) - T(m_1 - 1)\} = m_2 \{T(m_2) - T(m_2 - 1)\}, \tag{14}
\]
for all $m_1, m_2$ such that $m_1 + m_2 = R$. (Here we set $T(-1) = 0$). Any transform $T$ that satisfies the following:
\[
T(m_1) - T(m_1 - 1) = \frac{a}{m_1}, \quad m_1 = 1, 2, \cdots, R - 1, \text{ for some } a > 0, \text{ and } \\
T(R) = T(R - 1),
\]
should satisfy the equations in (14). Starting from an arbitrary initial value \( T(R) \), we can recursively recover the values of \( T(m_1) \) for each \( m_1 \in \{0, 1, 2, \cdots, R\} \). The resulting transform \( T \) takes the following form: for each \( m_1 = 0, 1, 2, \cdots, R-1 \),

\[
T(m_1) = - \sum_{s=0}^{R-m_1-1} \frac{a}{R-s} + T(R),
\]

where we take the summation \( \sum_{s=0}^{R-m_1-1} \) to be zero. By setting \( T(R) = 0 \) and \( a = 1 \), we obtain the formula in (7). For \( J \geq 3 \), the derivation of \( T \) from (14) follows similar arguments but is more involved. The solution for the case of a general \( J \) is given in the following algebraic result.

**Lemma 1:** Transform vector \( T = (T_1, \cdots, T_J) \) satisfies (12), if for each \( j = 1, \cdots, J \) and \( R = 2, 3, \cdots \), \( T_j(m) = T_{R,j}(m) \), where

\[
T_{R,j}(m) = - \sum_{s=0}^{R-m_j-1} \frac{1}{R-s} + \frac{\sum_{k=1, k\neq j}^{J} 1\{m_k > 0\}}{R}.
\] (15)

Lemma 1 is a pure algebraic result that does not involve any unknown specifics of the data generating process in the model. The condition (12) is only a necessary condition for (8). The sufficient second order condition for the optimization problem in (8) stems from the result of Lemma 2 below. Given \( p \in S_J \), we write \( \tilde{p} = (p_1, \cdots, p_{J-1})^\top \). We define

\[
\tilde{\Lambda}_R(\tilde{p}, p_0; T_R) = \Lambda_R((p_1, \cdots, p_{J-1}, 1 - \Sigma_{j=1}^{J-1} p_j), p_0; T_R).
\]

Hence \( \tilde{\Lambda}_R(\tilde{p}, p_0; T_R) \) is \( \Lambda_R(p, p_0; T_R) \) with the imposition of the constraint \( \Sigma_{j=1}^{J-1} p_j = 1 \).

**Lemma 2:** Let \( T_R = (T_{R,1}, \cdots, T_{R,J}) \) with \( T_{R,j} \) given by (15). Then for any \( a = (a_1, \cdots, a_{J-1})^\top \in \mathbb{R}^{J-1} \),

\[
a^\top \left( \frac{\partial^2 \tilde{\Lambda}_R(\tilde{p}, p_0; T_R)}{\partial \tilde{p} \partial \tilde{p}^\top} \right) a \leq -2\varepsilon \left( \sum_{j=1}^{J-1} a_j^2 \right)
\]

where \( \varepsilon > 0 \) is such that \( p_{j0} \in [\varepsilon, 1 - \varepsilon] \), for all \( j = 1, \cdots, J \).

Lemma 2 says that the function \( \tilde{\Lambda}_R(\tilde{p}, p_0; T_R) \) for all \( \tilde{p} = (p_1, \cdots, p_{J-1})^\top \) with \( p \in S_J \) is globally strictly concave in \( \tilde{p} \) if \( p_0 \) is such that for some \( \varepsilon > 0 \), \( p_{j0} \in [\varepsilon, 1 - \varepsilon] \), for all \( j = 1, \cdots, J \). Combined with Lemma 1, the result of Lemma 2 shows that \( \Lambda_R(p, p_0; T_R) \) (under constraint \( \Sigma_{j=1}^{J-1} p_j = 1 \)) is uniquely maximized at \( p = p_0 \). Formal proofs of Lemmas 1 and 2 are algebraically involved and provided in the appendix.
It may not be immediately clear how the choice of (15) is related to MLE with sufficiently large $R$. To see this connection, note first that the simulated frequencies $m_{ij}(\theta)/R \to_p \pi_{ij}(\theta) \in (0, 1)$ with $R \to \infty$, by the law of large numbers, where $\pi_{ij}(\theta) = \pi_j(X_i; \theta)$. Also, note that

$$0 \leq \frac{1}{R} \sum_{k=1, k \neq j}^J 1\{m_k > 0\} \leq \frac{J-1}{R} \to 0$$

as $R \to \infty$. Finally, observe that for large $R$,

$$- \sum_{s=0}^{R-m_{ij}(\theta)-1} \frac{1}{R-s} \approx \log(\pi_{ij}(\theta)).$$

This latter convergence is immediate as the sum on the left-hand side is a Riemann lower sum of $-\int_{m_{ij}(\theta)/R}^{1} (1/x)dx$ and $m_{ij}(\theta)/R \approx \pi_{ij}(\theta)$ for large $R$. Therefore,

$$\mathbb{E}[D_{ij}T_{R,j}(m_{ij}(\theta))] \approx \mathbb{E}[D_{ij} \log(\pi_{ij}(\theta))],$$

for large $R$ under regularity conditions. Hence the TSF population objective function is close to that of MLE for large $R$.

### 3.3 Asymptotic Properties

In this section we investigate the asymptotic properties of the estimator $\hat{\theta}$ defined in (6). The asymptotic properties of $\hat{\theta}$ are developed for two separate cases: the case with $R$ fixed and the case with $R$ tending to infinity jointly with the sample size $n$. Let $X$ be the support of $X_i$. We introduce the following assumptions.

**Assumption 2**: (i) $\{(D_i, X_i, \eta_i)\}_{i=1}^n$ is i.i.d. from a common distribution.

(ii) For each $\bar{\theta} \in \Theta$, $\sup_{x \in X} \mathbb{E}[\sup_{\theta \in B(\bar{\theta}, \varepsilon)} |\delta_j(X_i, \eta_i; \bar{\theta}) - \delta_j(X_i, \eta_i; \bar{\theta})|^2 |X_i = x] \leq C \varepsilon$, for some $C > 0$.

(iii) $(X_i, \eta_i)_{i=1}^n$ and $(X_i, \eta_i^*)_{i=1}^n$ are distributionally identical.

(iv) For each $x$ in the support of $X_i$ and for each $j \in \{1, \cdots, J\}$, $\pi_j(x; \theta)$ is twice continuously differentiable at $\theta = \theta_0$ and for some $\delta > 0$,

$$\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| \frac{\partial p(X_i; \theta)}{\partial \theta} \right| ^2 \right] < \infty \text{ and } \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| \frac{\partial^2 p(X_i; \theta)}{\partial \theta \partial \theta^T} \right| ^2 \right] < \infty,$$

where $p(X_i; \theta) = (\pi_1(X_i; \theta), \cdots, \pi_J(X_i; \theta))^T$.

Condition (ii) controls the manner the random decision rule $\delta_j(X_i, \eta_i; \theta)$ depends on $\theta$ and
The condition requires that the decision rule $\delta$ is locally uniformly $L_2$-continuous in $\theta$ (e.g., Chen, Linton, and van Keilegom (2003)). This condition is a very useful high-level condition that can be used to establish the stochastic equicontinuity of an empirical process involving a discontinuous function, and flexibly admits a wide class of specifications of $\delta$. See Example 1 for lower level conditions in the case of a random utility framework. Condition (iii) is certainly satisfied when $X_i$ and $\eta_i$ are independent and $\eta_i^*$ are drawn i.i.d from $F$, as commonly assumed in the simulation literature. Condition (iv) is a standard assumption that is often used in the literature of discrete choice models.

**Example 1:** We consider a static random utility model. Suppose that the utility of agent $i$ with covariates $X_i$ and stochastic errors $\eta_i$ when she makes the $j$-th choice is given by $u_j(X_i, \eta_{ij}; \theta)$. Then she makes the $j$-th choice when

$$\Delta_j(X_i, \eta_i; \theta) = u_j(X_i, \eta_{ij}; \theta) - \max_{1 \leq k \leq J, k \neq j} u_k(X_i, \eta_{ik}; \theta)$$

is greater than zero. In this case, the decision rule $\delta_j$ is defined by

$$\delta_j(X_i, \eta_i; \theta) = 1 \{\Delta_j(X_i, \eta_i; \theta) > 0\}.$$

Suppose that for each $\theta \in \Theta$, and for each $x$ in the support of $X$, and for $j = 1, 2, \cdots, J$,

$$\left| u_j(X_i, \eta_{ij}; \theta) - u_j(X_i, \eta_{ij}; \bar{\theta}) \right| \leq C_j(X_i, \eta_{ij})||\theta - \bar{\theta}||,$$

where $C_j(X_i, \eta_{ij})$ is such that $\mathbb{E} [C_j(X_i, \eta_{ij})^2] < \infty$. Furthermore, assume that the conditional density of $C_j(X_i, \eta_{ij}) - C_k(X_i, \eta_{ik})$ given $X_i = x$ is bounded uniformly over $x$ in the support of $X_i$. Then the condition of Assumption 2(ii) holds. ■

The following theorem establishes the rate of convergence for the TSF-MLE estimator $\hat{\theta}$ for finite $R$.

**Theorem 2 (The Rate of Convergence for Fixed $R$):** Suppose that Assumptions 1-2 hold. Then for each fixed $R \geq 2$, we have

$$n^{1/3}(\hat{\theta} - \theta_0) = O_P(1).$$

The estimator $\hat{\theta}$ follows the cube-root asymptotics of Kim and Pollard (1990) with fixed $R$ and $n \to \infty$. The cube-root asymptotics occur precisely in the same way it occurs in maximum score estimation. When $R$ tends to infinity slightly faster than $\sqrt{n}$, not only is the $\sqrt{n}$-rate of convergence restored, but also the estimator achieves the efficiency of MLE.
Theorem 3 (Asymptotic Normality): Suppose that Assumptions 1-2 hold. As \( n, R \to \infty \) jointly, with \( \sqrt{n} \log(R)/R \to 0 \),

\[
\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, V),
\]

where \( V = \Omega^{-1} \) and

\[
\Omega = \mathbb{E} \left[ \left( \sum_{j=1}^J D_{ij} \frac{\partial}{\partial \theta} \log p_j(X_i, \theta_0) \right) \left( \sum_{j=1}^J D_{ij} \frac{\partial}{\partial \theta} \log p_j(X_i, \theta_0) \right)^\top \right].
\]

The rate condition \( \sqrt{n} \log(R)/R \to 0 \) is satisfied when \( R \) increases slightly faster than \( \sqrt{n} \). This condition is nearly close to the usual condition \( (\sqrt{n}/R \to \infty) \) in the simulated likelihood literature. (See Gouriéroux and Monfort (1991)).

As for standard errors, we suggest using a consistent estimator of the asymptotic covariance matrix in Theorem 3 based on large \( R \) asymptotics. It suffices for the computation of its standard error to evaluate the asymptotic covariance matrix formula once and for all, whereas the estimation of \( \hat{\theta} \) requires evaluation of the sample objective function for each optimization iteration.

Instead of computing numerical derivatives of the choice probability based on the simulated frequencies, we suggest using numerical derivatives of TSF: \( T_{R,j}(m_i(\theta)) \) in \( \theta \). Define \( \varepsilon_n \) to be a sequence such that \( \varepsilon_n \to 0 \) and \( \varepsilon_n R \to \infty \). Then, we take \( \hat{g}_i \) to be the \( d \times 1 \) vector whose \( k \)-th entry is given by

\[
\frac{1}{\varepsilon_n} \sum_{j=1}^J D_{ij} \{ T_{R,j}(m_i(\hat{\theta} + \varepsilon_ne_k)) - T_{R,j}(m_i(\hat{\theta})) \},
\]

where \( e_k \) denotes a unit vector whose \( k \)-th entry is one and the other entries are zero. Using numerical derivatives of a discontinuous finite sample function to compute standard errors is not new here (e.g. see Sherman (1993)). Then we define

\[
S_R = \frac{1}{n} \sum_{i=1}^n \hat{g}_i \hat{g}_i^\top.
\]

The standard error for the \( j \)-th parameter \( \theta_{0,j} \) in \( \theta_0 \) is taken to be the square-root of the \( j \)-th diagonal element of \( S_R^{-1} \). The covariance matrix estimator is consistent as \( n, R \to \infty \). In the simulation studies, we compare the performance of asymptotic approximation using various different covariance matrix estimators.
Example 2: Cohort-Level Aggregate Data: Suppose that we have $K$ number of cohorts and $n(k)$ number of agents in the $k$-th cohort. The individual decision variable $D_{ij}(k)$ corresponding to the agent $i$ in cohort $k$ choosing the $j$-th choice is defined as a binary variable such that

$$D_{ij}(k) = \delta_j(X(k), \eta_{ij}(k); \theta),$$

when the $j$-th choice is made by the agent $i$ in cohort $k$.

Note that the observed variable $X(k)$ is only a cohort-level aggregate covariate. The variables $D_{ij}(k)$ and $\eta_{ij}(k)$ represent the unobserved micro variables for each individual. Define

$$D_j(k) = \frac{1}{n(k)} \sum_{i=1}^{n(k)} D_{ij}(k)$$

and $D(k) = (D_1(k), \cdots, D_J(k))^\top$. The variable $D_j(k)$ indicates a proportion of agents in cohort $k$ that have chosen the $j$-th choice. The econometrician observes only the cohort-level aggregate data $\{D(k), X(k)\}_{k=1}^{K}$. The (infeasible) log-likelihood of the micro data after normalizing by $n(k)$ is equal to

$$\sum_{k=1}^{K} \sum_{j=1}^{J} \frac{1}{n(k)} \sum_{i=1}^{n(k)} D_{ij}(k) \log \mathbb{P} \{D_{ij}(k) = 1|X(k), \theta\}.$$

When the conditional distribution of the stochastic error $\eta_{ij}(k)$ given $X(k)$ is identical for each individual $i$, the conditional probability $\mathbb{P} \{D_{ij}(k) = 1|X(k), \theta\}$ is identical for all the individuals in the $k$-th cohort. This is the case when $\{\eta_{ij}(k) : i = 1, \cdots, n(k), \ k = 1, \cdots, K\}$ is i.i.d. and independent of $\{X(k) : k = 1, \cdots, K\}$. In this case, we can write the cohort-level likelihood as

$$\sum_{k=1}^{K} \sum_{j=1}^{J} D_j(k) \log \mathbb{P} \{D_{ij}(k) = 1|X(k), \theta\}.$$

This is the log-likelihood using only the observable cohort characteristics and the proportion of agents in each cohort that made certain decisions. Let $F$ be the fully known marginal distribution of $(\eta_{i1}(k), \cdots, \eta_{iJ}(k))$. Then, one draws $R$ random sample from $F$ to obtain $\{\eta_{r}(k)\}_{r=1}^{R} \eta_{rJ}(k))$. We define the simulated frequency

$$m_{jr}(k, \theta) = \frac{1}{R} \sum_{r=1}^{R} \delta_j(X(k), \eta_{rJ}(k); \theta).$$

Then using the transform that we propose here, we can construct an objective function as
follows
\[ l_{K,R}^*(\theta) = \frac{1}{K} \sum_{k=1}^{K} \sum_{j=1}^{J} D_j(k)TR_{Rj}(m_{R}(k, \theta)). \]

Note that
\[ E[D_j(k)|X(k)] = P\{D_{ij}(k) = 1|X(k)\}. \]

Hence one can check sufficient conditions with this choice probability. The results of Theorems 1-3 carry over to this case as long as the data \{D(k), X(k)\}_{k=1}^{K} are cohort-wise i.i.d.

4 Monte Carlo Study

We performed a Monte Carlo simulation study based on three different models. The first model is a simple multinomial logit model where we can evaluate the choice probability explicitly and compare simulation-based estimators with MLE. The second model is a dynamic schooling choice model where the unobserved heterogeneity in the payoff functions is time invariant. In this case, the model can be written as a static discrete choice model. The third model is a dynamic schooling model where unobserved heterogeneity in the payoff functions is time-varying. In this case, one cannot reduce the dynamic model to a static discrete choice model.

4.1 Multinomial Logit Models

We consider the following standard logit models, where we can explicitly compare the true MLE, the Lerman-Manski simulated frequency method, and the transformed simulated frequency method. The choice probability is specified as follows:

\[ P\{D_{i1} = 1|X_i\} = \frac{1}{1 + \exp(X_{i1}\beta_1) + \exp(X_{i2}\beta_2)}, \]
\[ P\{D_{i2} = 1|X_i\} = \frac{\exp(X_{i1}\beta_1)}{1 + \exp(X_{i1}\beta_1) + \exp(X_{i2}\beta_2)}, \text{ and} \]
\[ P\{D_{i3} = 1|X_i\} = \frac{\exp(X_{i2}\beta_2)}{1 + \exp(X_{i1}\beta_1) + \exp(X_{i2}\beta_2)}. \]

As for the distribution of \((X_{i1}, X_{i2})^T\), the study considered the following three different specifications. Let \(V_i \sim \text{Uniform}[0, 1]\), \(Z_i \sim N(0, 1)\), \(B_{i1} \sim \text{Binomial}(2, 0.3)\), \(B_{i2} \sim \text{Binomial}(2, 0.5)\), and \(W_i = Z_{i1} + (B_{3i} - 1/2)/4\), where \(B_{3i} \sim \text{Bernoulli}(0.3)\) and \(Z_{i1} \sim N(0, 1)\). The random variables \(V_i, Z_i, Z_{i1}, B_{i1}, B_{i2}\), and \(B_{3i}\) were drawn independently. Using these random vari-
ables, we specified $X_{1i}$ and $X_{2i}$ as follows:

- **Specification A:**
  \[
  X_{1i} = Z_{1i} + V_i \\
  X_{2i} = Z_{2i} + V_i
  \]

- **Specification B:**
  \[
  X_{1i} = \Phi(Z_{1i} + V_i) - 1 + V_i \\
  X_{2i} = 2U_{1i} - 4U_{2i}^2 + V_i
  \]

- **Specification C:**
  \[
  X_{1i} = \Phi(Z_{1i}/2 + (B_{1i}/2 - 1)/4) - 1/2 + W_i \\
  X_{2i} = Z_{i}/2 + \Phi(B_{2i}/2 - 1)/4 + W_i - 1/2.
  \]

The Monte Carlo simulation number was taken to be 5000.

Figure 1 reports the average of the mean absolute deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$. Both in the cases of the sample size equal to 300 and 1000, the TSF dominates the Lerman-Manski’s simulated frequency method when the simulation number is small ranging from 10 to 100. When the sample size is 300, this order of dominance is slightly reversed when the simulation number is larger, although both converging to the mean absolute deviation of the MLE. However, when the sample size is 1000, the estimator from the TSF method dominates that of the Lerman-Manski’s method uniformly over all the simulation numbers considered. As expected, the dominance is prominent when the simulation number is small. This is because the TSF method delivers a consistent estimator even for a small simulation number while the Lerman-Manski’s simulated frequency method does not.

### 4.1.1 Finite Sample Coverage Probabilities of Confidence Intervals

Here we report the performance of the 95% confidence intervals for the parameters $\beta_2$ and $\beta_3$. Since we used a numerical derivative method as is usually done in the case of a non-smooth objective function (e.g. Sherman (1993)), it is important to set the step size appropriately for the numerical derivative. We present results from using the step size of $\varepsilon_n = c \in \{0.6, 0.8, 1.0\}$. We also report only the coverage probabilities computed using the TSF method. The performance of the Lerman-Manski’s simulated frequency method was similar, only requiring a slightly different range of $c$’s to achieve coverage probabilities close to 95%.
Figure 3: Average Mean Absolute Deviation of Estimators of $\beta_2$ and $\beta_3$
Table 2: Finite Sample Coverage Probabilities of 95% Confidence Intervals (Specification A)

<table>
<thead>
<tr>
<th>$n = 500$</th>
<th>$R = 200$</th>
<th>$R = 400$</th>
<th>$R = 600$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = 0.6$</td>
<td>0.9142</td>
<td>0.9508</td>
<td>0.9660</td>
</tr>
<tr>
<td>$c = 0.8$</td>
<td>0.9678</td>
<td>0.9848</td>
<td>0.9910</td>
</tr>
<tr>
<td>$c = 1.0$</td>
<td>0.9798</td>
<td>0.9934</td>
<td>0.9930</td>
</tr>
</tbody>
</table>

Table 3: Finite Sample Coverage Probabilities of 95% Confidence Intervals (Specification B)

<table>
<thead>
<tr>
<th>$n = 500$</th>
<th>$R = 200$</th>
<th>$R = 400$</th>
<th>$R = 600$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = 0.6$</td>
<td>0.8680</td>
<td>0.8956</td>
<td>0.9084</td>
</tr>
<tr>
<td>$c = 0.8$</td>
<td>0.9376</td>
<td>0.9474</td>
<td>0.9544</td>
</tr>
<tr>
<td>$c = 1.0$</td>
<td>0.9654</td>
<td>0.9798</td>
<td>0.9840</td>
</tr>
</tbody>
</table>

Table 4: Finite Sample Coverage Probabilities of 95% Confidence Intervals (Specification C)

<table>
<thead>
<tr>
<th>$n = 500$</th>
<th>$R = 200$</th>
<th>$R = 400$</th>
<th>$R = 600$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = 0.6$</td>
<td>0.8922</td>
<td>0.9030</td>
<td>0.9048</td>
</tr>
<tr>
<td>$c = 0.8$</td>
<td>0.9420</td>
<td>0.9516</td>
<td>0.9542</td>
</tr>
<tr>
<td>$c = 1.0$</td>
<td>0.9674</td>
<td>0.9736</td>
<td>0.9740</td>
</tr>
</tbody>
</table>

The results are presented in Tables 2-4. The results show that the performance of the finite sample coverage probability becomes better with a wider range of $c$’s as we increase the simulation number. Nevertheless, the finite sample coverage probabilities are sensitive to the choice of the step size $c$ as we expected. This issue applies in general to the method of using a numerical derivative in a situation involving nonsmooth simulated objective functions.

4.2 Schooling Choice with Time Invariant Unobserved Heterogeneity

4.2.1 The Data Generating Process

In this section, we present and discuss results from a Monte Carlo simulation study based on a model of schooling choice. The model involves observed ability affecting each agent’s labor market outcome, and unobserved heterogeneity in discount factor and preference. See Willis and Rosen (1979) and Keane and Wolpin (1997) for structural models of education with unobserved heterogeneity in the preferences.
Suppose that people make schooling decisions at the age 16 endowed with 10 years of education. They can choose among the 4 alternatives: 1) to drop out of high school and start working right away, 2) to graduate from high school attaining 12 years of education, 3) to graduate a 2-year college with 14 years of education, and 4) to graduate from college with 16 years of education. After finishing their respective schooling, they work until age 65 and there is no labor supply decision. Therefore, the number of periods in the model is 50 periods.

People are assumed to be heterogeneous in 1) two observed measures of ability ($X_1$ and $X_2$) which affect their labor market income, 2) unobserved discount factor and 3) unobserved random utility value of schooling. Labor market income is determined by individuals' ability and years of schooling and is assumed to follow the Mincer-type specification:

$$w_t = \exp \left( \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 E + \varepsilon_w \right),$$

where $E$ is the years of education taking values of 10, 12, 14, and 16 and $\varepsilon_w$ is normal, i.i.d., across individuals and periods with standard deviation of $\sigma_w$. Once an individual enters the labor market and starts working, going back to school is not permitted in the model.

In each period $t$, the utility is given by $U_{w,t}$ if the individual works, and $U_{s,t}$ if he attends school. Also, we assume that the individual observes the labor income shock only after he enters the labor market and, therefore, the expected value of the wage only enters the utility function. This set-up yields the following two utilities from entering the labor market and from school:

$$U_{w,t} = \mathbb{E}(w_t|X_1, X_2, E_t) = \exp \left( \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 E_t + \frac{1}{2} \sigma^2 \right),$$

$$U_{s,t} = \gamma_1 1\{\text{in high school}\} + \gamma_2 1\{\text{in two-year college}\} + \gamma_3 1\{\text{in four-year college}\} + \varepsilon_s,$$

where $E_t$ denotes the years of education received up to $t$, so that

$$E_{t+1} = E_t + 1\{\text{schooling is chosen at } t\}.$$

Here $\gamma_1$ is the average utility of attending high school (we assume that there is no tuition for attending high school), $\gamma_2$ the average utility of attending two year college including tuition cost, $\gamma_3$ the average utility of attending four year college including tuition cost, $\varepsilon_s$ is mean zero and normally distributed individual specific random effect on schooling utility which is independent across individuals, but is fixed over time for each individual. The standard deviation of $\varepsilon_s$ is denoted by $\sigma_s$.

We assume that the discount factor $\delta$ is heterogeneous across people and is correlated
with an observed covariate $X_3$. It is specified as

$$\delta = \frac{1}{1 + \exp(\varepsilon_\delta + \rho_0 + \rho_1 X_3)},$$

(16)

where $\varepsilon_\delta$ is normally distributed with mean 0 and standard deviation $\sigma_\delta$ and does not change over time for each individual. The errors $\varepsilon_w, \varepsilon_s$, and $\varepsilon_\delta$ are independent.

Let $E$ be the total amount of schooling and $U(E = a)$ be the discounted utility from schooling choice $E = a$ at the beginning of one’s lifecycle. Given that working is an absorbing state, we can represent this multi period dynamic programming model as a 4-choice static model with the following random utilities:

$$\begin{align*}
U(E = 10) &= \sum_{t=1}^{50} \delta^{t-1} e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + 10 + \frac{1}{2} \sigma_w^2} \\
U(E = 12) &= \sum_{t=1}^{2} \delta^{t-1} (\gamma_1 + \varepsilon_a) + \sum_{t=3}^{50} \delta^{t-1} e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + 12 + \frac{1}{2} \sigma_w^2} \\
U(E = 14) &= \sum_{t=1}^{2} \delta^{t-1} (\gamma_1 + \varepsilon_a) + \sum_{t=3}^{4} \delta^{t-1} (\gamma_2 + \varepsilon_a) + \sum_{t=5}^{50} \delta^{t-1} e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + 14 + \frac{1}{2} \sigma_w^2} \\
U(E = 16) &= \sum_{t=1}^{2} \delta^{t-1} (\gamma_1 + \varepsilon_a) + \sum_{t=3}^{6} \delta^{t-1} (\gamma_3 + \varepsilon_a) + \sum_{t=7}^{50} \delta^{t-1} e^{\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + 16 + \frac{1}{2} \sigma_w^2}.
\end{align*}$$

Given the model structure, we expect people with higher ability $X_1$ and $X_2$, higher discount factor $\delta$ and higher utility value of schooling $\varepsilon_s$ to attain a higher level of schooling.

We assume that the econometrician observes the ability measures $X_1$ and $X_2$, the schooling outcome, and characteristics $X_3$ that affect discount factor. Discount factor $\delta$ and the utility value of schooling $\varepsilon_2$ are not observed. For simplicity, we assume that the parameters in the wage equation are known and focus only on the parameters in the schooling utility and the parameters in the discount factor. Hence the parameters of interest in this exercise are as follows:

- schooling utility parameters: $\gamma_1, \gamma_2, \gamma_3, \sigma_s$
- discount factor parameters: $\rho_0, \rho_1, \sigma_\delta$. 

23
For comparison with our TSF-MLE, we considered MSL estimator following Lerman and Manski (1981)’s simulated frequency method, and its smoothed version (McFadden (1989)). The Lerman and Manski ’s simulated frequency method uses simulated frequencies to compute simulated choice probabilities. To prevent the zero-probability problem, we substituted $0.5/R$ for simulated probabilities that turned out to be zero. The second kind is a smoothed MSL estimator which is computed by using the following smoothed simulated choice probability:

$$\frac{1}{R} \sum_{r=1}^{R} \frac{\exp(U_{j,r,t}/\lambda)}{\sum_{j=1}^{J} \exp(U_{j,r,t}/\lambda)}.$$  \hspace{1cm} (17)

Here the parameter $\lambda$ is a smoothing parameter, larger values indicating more smoothing, and $U_{j,r,t}$ denotes the simulated value function at $t$ of choice $j$ at the $r$-th simulation. The smoothing parameter chosen from $\{0.1, 0.01\}$ performed relatively better than other choices.

Note that except for the simulated frequency method of Lerman and Manski (1981) or its smoothed version, we cannot apply the existing simulation methods that require the presence of additive normal or logit errors in the random utility due to nonlinear unobserved heterogeneity in discount factor.

The sample size was chosen among $\{100, 200, 500, 1000\}$ and the simulation number from $\{10, 20, 50, 100\}$. When the simulation number was equal to or greater than 100, the comparison was not much informative as most estimators perform well in our data generating process. The Monte-Carlo simulation number was set to be 1000. The parameter values are chosen as follows. As for wage parameters, $\alpha_1 = 8$, $\alpha_2 = 1$, $\alpha_3 = 1$, $\alpha_4 = 0.07$, and $\sigma_w = 0.3$. As for the schooling utilities, $\gamma_1 = 0$, $\gamma_2 = -5000$, $\gamma_3 = -20,000$, and $\sigma_s = 5000$. And finally, as for discount factors, $\rho_0 = -0.25$ and $\rho_1 = 0.2$.

### 4.2.2 The Results

The results are reported in Tables 5-8. In Table 5 we compare the overall simulation errors in terms of the log-likelihood evaluation of the simulation-based estimator using the true log-likelihood $l_n(\theta)$. This number is bounded by $l_n(\hat{\theta}_{MLE})$ with $\hat{\theta}_{MLE}$ denoting the MLE of $\theta_0$. As the number is higher, the simulation-based estimator suffers from a smaller overall simulation error. First, note that the performance of the Lerman-Manski’s simulated frequency method is different from its smoothed version. The simulation results show that the use of smoothing does not necessarily improve the performance, and sometimes, even worsen the quality of the estimator.
### Table 5: True Log Likelihood Evaluated at the Estimators

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>-857.99</td>
<td>-853.07</td>
<td>-848.21</td>
<td>-845.89</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>-912.13</td>
<td>-859.81</td>
<td>-848.93</td>
<td>-849.82</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.1$)</td>
<td>-905.51</td>
<td>-871.45</td>
<td>-849.19</td>
<td>-845.93</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.01$)</td>
<td>-903.98</td>
<td>-871.36</td>
<td>-849.19</td>
<td>-846.00</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>-855.51</td>
<td>-852.52</td>
<td>-850.57</td>
<td>-849.78</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>-907.59</td>
<td>-858.35</td>
<td>-851.15</td>
<td>-850.81</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.1$)</td>
<td>-937.18</td>
<td>-906.90</td>
<td>-856.37</td>
<td>-850.81</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.01$)</td>
<td>-935.56</td>
<td>-906.91</td>
<td>-856.39</td>
<td>-850.85</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>-853.35</td>
<td>-852.44</td>
<td>-851.85</td>
<td>-851.57</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>-905.72</td>
<td>-856.25</td>
<td>-852.12</td>
<td>-851.58</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.1$)</td>
<td>-962.27</td>
<td>-949.84</td>
<td>-872.51</td>
<td>-855.93</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.01$)</td>
<td>-960.47</td>
<td>-948.46</td>
<td>-872.59</td>
<td>-855.99</td>
</tr>
</tbody>
</table>

### Table 6: MAE of Estimated $\gamma_1$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>0.038</td>
<td>0.036</td>
<td>0.031</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.086</td>
<td>0.041</td>
<td>0.030</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.1$)</td>
<td>0.068</td>
<td>0.043</td>
<td>0.027</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.01$)</td>
<td>0.067</td>
<td>0.042</td>
<td>0.027</td>
<td>0.029</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>0.026</td>
<td>0.023</td>
<td>0.020</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.080</td>
<td>0.029</td>
<td>0.019</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.1$)</td>
<td>0.098</td>
<td>0.083</td>
<td>0.023</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.01$)</td>
<td>0.097</td>
<td>0.082</td>
<td>0.024</td>
<td>0.017</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>0.016</td>
<td>0.015</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.086</td>
<td>0.021</td>
<td>0.011</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.1$)</td>
<td>0.121</td>
<td>0.123</td>
<td>0.046</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski ($\lambda = 0.01$)</td>
<td>0.121</td>
<td>0.121</td>
<td>0.047</td>
<td>0.017</td>
</tr>
</tbody>
</table>
When the sample size is small, the performance of Lerman-Manski’s simulated frequency method and its smoothed version becomes comparable to our methods. However, when the sample size is large, the improved performance of our estimator becomes prominent over that of the competing procedures. This confirms our theoretical result that our estimator is consistent even when the simulation number is small, but the Lerman-Manski’s procedures do not possess this property.
A similar pattern of performance comparison is obtained in terms of mean absolute errors (MAE) of individual estimators, as shown in Tables 6-8. While not reported here, we observed a similar pattern of performance for other parameters.

Lastly, in Table 9, we report a sample of computing time for each method. Overall, it is easily seen that as one increases \( n \) and \( R \), the computation time increases. This suggests that an estimator that maintains good quality for a smaller simulation number \( R \) is also computationally more convenient. The computing times for the TSF method and Lerman-Manski’s simulated frequency method turned out to be similar, except when \( R = 100 \) and \( n = 100 \) or 200. In our simulation study, the smoothed version of Lerman-Manski’s method does not show computational efficiency.

Table 9: Computing Time for Obtaining a Point Estimate (in Median Seconds from 1000 Simulations)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>( R = 10 )</th>
<th>( R = 20 )</th>
<th>( R = 50 )</th>
<th>( R = 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 100 )</td>
<td>TSF-MLE</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski (( \lambda = 0.1 ))</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski (( \lambda = 0.01 ))</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>27</td>
</tr>
<tr>
<td>( n = 200 )</td>
<td>TSF-MLE</td>
<td>4</td>
<td>11</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski (( \lambda = 0.1 ))</td>
<td>7</td>
<td>12</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski (( \lambda = 0.01 ))</td>
<td>8</td>
<td>15</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>( n = 500 )</td>
<td>TSF-MLE</td>
<td>9</td>
<td>9</td>
<td>17</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>10</td>
<td>14</td>
<td>16</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski (( \lambda = 0.1 ))</td>
<td>18</td>
<td>16</td>
<td>26</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski (( \lambda = 0.01 ))</td>
<td>20</td>
<td>24</td>
<td>25</td>
<td>54</td>
</tr>
<tr>
<td>( n = 1000 )</td>
<td>TSF-MLE</td>
<td>12</td>
<td>15</td>
<td>32</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>13</td>
<td>15</td>
<td>30</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski (( \lambda = 0.1 ))</td>
<td>20</td>
<td>28</td>
<td>57</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Smoothed Lerman-Manski (( \lambda = 0.01 ))</td>
<td>32</td>
<td>43</td>
<td>56</td>
<td>109</td>
</tr>
</tbody>
</table>
4.3 Schooling Choice with Time-Varying Unobserved Heterogeneity

4.3.1 The Data Generating Process

The schooling choice model considered here is the same as the previous model except that the utility of schooling involves an idiosyncratic shock each period which is observed by the individual before he makes a choice but is not observed by the econometrician. More specifically, the utility from schooling at period $t$ is given by

$$U_{st}(E_t) = \gamma_0 + \gamma_1 1\{E_t \geq 4\} + \varepsilon_{s,t},$$

where $\varepsilon_{s,t}$ is the idiosyncratic shock. Here $\gamma_0$ is utility of schooling in the first 4 years of schooling and $\gamma_1$ is the additional utility of schooling, potentially associated with the tuition cost of college in which case $\gamma_1 < 0$.

The decision rule for an individual is written in a standard dynamic programming framework. Given that leaving schooling is an absorbing state, we look at the decisions at $t$ who have continued to school up to $t - 1$ with $E = t - 1$. The value of attending school and the value of working at education level $E$ and at period $t$ are given by

$$V_s(t, E) = U_{st}(E) + \delta\mathbb{E} \left[ \max_{1 \leq \tau \leq t} \{V_s(t + 1, E + 1), V_w(E + 1)\} \mid X, E, \varepsilon \right],$$

$$V_w(t, E) = U_{w,t}(E) + \sum_{\tau=1}^{T-t} \delta^\tau U_{w,t}(E),$$

where $\varepsilon_\delta$ is the error term in (16) and $X = (X_1, X_2, X_3)^T$. Here

$$U_{w,t}(E) = \mathbb{E}(w_t|X_1, X_2, E) = \exp \left( \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 E + \frac{1}{2} \sigma_w^2 \right).$$

Individuals will attend school if $V_s(t, E) > V_w(t, E)$.

While the expected value function $V_s(t, E)$ depends not only on the observable variables of $X$’s, but also on the unobservable component of $\eta$, the choice probability of education requires integration over $\eta$. In this situation, the method of smoothed simulated frequencies such as (17) is quite cumbersome, because one then needs to smooth choice probabilities of the binary decision on schooling or work from $t = 0$ to $t = E$, as well as the probability of leaving school at $t = E + 1$. On the other hand, the method of TSF and Lerman-Manski’s simulated frequencies is computationally efficient, because one can directly count the number of simulated outcomes that match the given amount of schooling observed for each individual.

The simulation was performed as follows. First, we drew realizations from the distrib-
utions of $\eta$ and $X$ that were set to be time-invariant. Then, for each simulated period, we
drew realizations from the distribution of $\varepsilon_{s,t}$, and calculated simulated versions of $V_s(t, E)$
and $V_w(t, E)$. These simulated versions of value functions as well as the given levels of $X$’s
and $\eta$ constitute individual decision rules. Following these decision rules, we simulated in-
dividual work/schooling choices and saved the results. We performed the same steps now
beginning with another set of realizations from the distributions of $\eta$ and $X$. By repeating
this process, we obtained simulated frequencies of each individual’s choice corresponding to
different values of $X$’s and $\eta$. Using these simulated frequencies, we performed simulated
maximum likelihood estimation in two different ways: one using our TSF-based method and
the other following Lerman-Manski’s simulated frequency method. Note that there are 11
discrete choices ($E = 0$ to $E = 10$) which come from the binary choices between work and
school at each point of time over the lifecycle. (Recall that in this model, once an individual
decides to work, she cannot come back to school for the rest of her life.)

The parameters used in the simulation study are as follows. As for schooling parameters,$\gamma_0 = 10,000, \gamma_1 = -10,000$, and $\sigma_s = 5,000$. Regarding observable characteristics vector
$X$, it is specified as follows.

$$
\begin{pmatrix}
    X_1 \\
    X_2 \\
    X_3 
\end{pmatrix} = 
\begin{pmatrix}
    1 & 0 & 0 \\
    \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\
    \frac{1}{2} & -\frac{1}{2} & \frac{1}{\sqrt{2}} 
\end{pmatrix}
\begin{pmatrix}
    \nu_1 \\
    \nu_2 \\
    \nu_3 
\end{pmatrix},
$$

where $\nu$’s are i.i.d. standard normal.

In the actual simulation, we fix all the wage parameters, and estimate the six of the
schooling parameters and discount factor parameters. The likelihood values based on the
estimators and the mean absolute deviations from the true parameters are reported.

### 4.3.2 The Results

The results are reported in Tables 10-16. The performance of TSF-MLE overall performs
better than that of Lerman-Manski’s simulated frequency method. In terms of true log-
likelihood evaluated at the estimators, the estimator based on TSF-MLE is closer to the
true MLE than the estimator of Lerman-Manski’s simulated frequency method (Table 10).
Also the MAEs of parameter estimates from TSF-MLE are mostly smaller than those from
Lerman-Manski’s simulated frequency method.
Table 10: True Log Likelihood Evaluated at the Estimators

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>-1695.69</td>
<td>-1690.51</td>
<td>-1685.33</td>
<td>-1682.72</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>-1829.93</td>
<td>-1711.97</td>
<td>-1686.86</td>
<td>-1686.24</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>-1692.47</td>
<td>-1690.45</td>
<td>-1687.17</td>
<td>-1685.97</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>-1837.39</td>
<td>-1713.67</td>
<td>-1688.82</td>
<td>-1686.24</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>-1689.55</td>
<td>-1688.54</td>
<td>-1687.55</td>
<td>-1687.12</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>-1840.09</td>
<td>-1713.68</td>
<td>-1689.38</td>
<td>-1687.36</td>
</tr>
</tbody>
</table>

Table 11: MAE of Estimated $\rho_0$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>0.0307</td>
<td>0.0283</td>
<td>0.0249</td>
<td>0.0234</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.0580</td>
<td>0.0396</td>
<td>0.0274</td>
<td>0.0235</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>0.0198</td>
<td>0.0184</td>
<td>0.0159</td>
<td>0.0156</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.0562</td>
<td>0.0344</td>
<td>0.0208</td>
<td>0.0159</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>0.0124</td>
<td>0.0111</td>
<td>0.0099</td>
<td>0.0088</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.0600</td>
<td>0.0341</td>
<td>0.0169</td>
<td>0.0107</td>
</tr>
</tbody>
</table>

Table 12: MAE of Estimated $\rho_1$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>0.0331</td>
<td>0.0304</td>
<td>0.0278</td>
<td>0.0243</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.0959</td>
<td>0.0548</td>
<td>0.0326</td>
<td>0.0249</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>0.0207</td>
<td>0.0197</td>
<td>0.0167</td>
<td>0.0157</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.1042</td>
<td>0.0513</td>
<td>0.0231</td>
<td>0.0167</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>0.0132</td>
<td>0.0114</td>
<td>0.0099</td>
<td>0.0088</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.1087</td>
<td>0.0482</td>
<td>0.0170</td>
<td>0.0104</td>
</tr>
</tbody>
</table>

Table 13: MAE of Estimated $\sigma_{\delta}$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>0.0107</td>
<td>0.0107</td>
<td>0.0105</td>
<td>0.0102</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.0110</td>
<td>0.0097</td>
<td>0.0095</td>
<td>0.0098</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>0.0097</td>
<td>0.0099</td>
<td>0.0091</td>
<td>0.0095</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.0093</td>
<td>0.0079</td>
<td>0.0078</td>
<td>0.0078</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>0.0090</td>
<td>0.0085</td>
<td>0.0080</td>
<td>0.0075</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>0.0083</td>
<td>0.0068</td>
<td>0.0057</td>
<td>0.0058</td>
</tr>
</tbody>
</table>
Table 14: MAE of Estimated $\gamma_0$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>286.15</td>
<td>269.50</td>
<td>251.72</td>
<td>241.27</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>908.56</td>
<td>378.80</td>
<td>298.93</td>
<td>268.83</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>228.25</td>
<td>220.92</td>
<td>201.30</td>
<td>200.21</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>1159.10</td>
<td>347.30</td>
<td>287.80</td>
<td>239.13</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>161.32</td>
<td>156.11</td>
<td>151.45</td>
<td>144.59</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>1652.87</td>
<td>406.66</td>
<td>305.05</td>
<td>219.30</td>
</tr>
</tbody>
</table>

Table 15: MAE of Estimated $\gamma_1$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>279.91</td>
<td>242.48</td>
<td>221.05</td>
<td>222.13</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>1427.69</td>
<td>471.45</td>
<td>251.70</td>
<td>232.17</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>210.76</td>
<td>194.12</td>
<td>179.55</td>
<td>179.12</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>1632.24</td>
<td>450.25</td>
<td>229.41</td>
<td>192.51</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>145.05</td>
<td>135.32</td>
<td>139.53</td>
<td>131.75</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>2094.36</td>
<td>497.22</td>
<td>237.65</td>
<td>168.65</td>
</tr>
</tbody>
</table>

Table 16: MAE of Estimated $\sigma_s$

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>$R = 10$</th>
<th>$R = 20$</th>
<th>$R = 50$</th>
<th>$R = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 300$</td>
<td>TSF-MLE</td>
<td>823.64</td>
<td>764.97</td>
<td>645.38</td>
<td>592.98</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>1493.10</td>
<td>839.97</td>
<td>569.71</td>
<td>560.06</td>
</tr>
<tr>
<td>$n = 1000$</td>
<td>TSF-MLE</td>
<td>633.43</td>
<td>579.57</td>
<td>448.05</td>
<td>390.52</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>1709.34</td>
<td>863.23</td>
<td>382.59</td>
<td>371.73</td>
</tr>
<tr>
<td>$n = 5000$</td>
<td>TSF-MLE</td>
<td>398.67</td>
<td>337.53</td>
<td>262.41</td>
<td>211.04</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>1797.97</td>
<td>961.55</td>
<td>256.57</td>
<td>192.67</td>
</tr>
</tbody>
</table>

In Table 17, we report the computing time taken for this simulation study. The computing time is substantially longer than that from the previous dynamic schooling model. This again emphasizes the fact that using a smaller simulation number in the case of a large sample size is computationally more convenient. The computing times for the TSF method and Lerman-Manski’s method are overall similar. This affirms our claim that using TSF does not cause much additional computational cost beyond that of Lerman-Manski.
Table 17: Computing Time for Obtaining a Point Estimate (in Seconds on Average from 1000 Simulations)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Simulation Methods</th>
<th>R = 10</th>
<th>R = 20</th>
<th>R = 50</th>
<th>R = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 300</td>
<td>TSF-MLE</td>
<td>85.6</td>
<td>107.9</td>
<td>193.6</td>
<td>306.3</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>102.7</td>
<td>134.7</td>
<td>242.4</td>
<td>302.2</td>
</tr>
<tr>
<td>n = 1000</td>
<td>TSF-MLE</td>
<td>199.4</td>
<td>252.9</td>
<td>438.0</td>
<td>759.1</td>
</tr>
<tr>
<td></td>
<td>Lerman-Manski</td>
<td>268.8</td>
<td>331.6</td>
<td>465.7</td>
<td>762.4</td>
</tr>
<tr>
<td>n = 5000</td>
<td>TSF-MLE</td>
<td>839.6</td>
<td>1066.8</td>
<td>1902.1</td>
<td>3296.2</td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper we propose an alternative method of MSL for discrete choice models that is applicable in various specifications of random utilities. While the method is as easy to apply as Lerman-Manski’s simulated frequency method, it is free from the issue of zero simulated choice probabilities and the issue of simulation bias. Furthermore, when the simulation number is large, the estimator is bound to achieve the efficiency of the infeasible MLE. This advantage is demonstrated both through the asymptotic result that shows consistency of the estimator with a finite simulation number and through various Monte Carlo simulation results.

6 Appendix: Proofs of the Results

Throughout the proofs, the notation $C$ denotes a constant that can take different values in different places.

Proof of Lemma 1: Let $T = (T_1, \cdots, T_J)$ be given collection of maps $T_j : \mathbb{N}_{R,J} \rightarrow \mathbb{R}$, $j = 1, \cdots, J$, such that for each $j = 1, 2, \cdots, J$,

$$T_j(m) = \tilde{T}(m_j, m_{-j}),$$

(18)

for a fixed map $\tilde{T} : \mathbb{N}_{R,J} \rightarrow \mathbb{R}$, where $m_{-j} = (m_1, \cdots, m_{j-1}, m_{j+1}, \cdots, m_J)$. Using this map, we write

$$A_R(p, p_0; T) = \sum_{j=1}^{J} p_{j_0} \sum_{m \in \mathbb{N}_{R,J}} T_j(m) \left( \begin{array}{c} R \\ m_{1,\cdots,m_{-j}} \end{array} \right) p_1^{m_1} \cdots p_J^{m_j}.$$ 

Note that the derivative of $A_R$ with respect to $p_k$ at $p = p_0$ is

$$\lambda_k(p_0, p_0; T) = \frac{\partial}{\partial p_k} A_R(p, p_0; T) \big|_{p = p_0}$$

(19)

$$= \sum_{j=1}^{J} p_{j_0} \sum_{m \in \mathbb{N}_{R,J}} \tilde{T}(m_j, m_{-j}) \left( \begin{array}{c} R \\ m_{1,\cdots,m_{-j}} \end{array} \right) m_k p_1^{m_1} p_2^{m_2} \cdots p_{k_0}^{m_{k_0}} \cdots p_{j_0}^{m_{j_0}}.$$
Let \( c_k(m_1, \cdots, m_J) \) be the coefficient of \( p_1^{m_1} \cdots p_J^{m_J} \) in the expansion of \( \lambda_k(p, p_0; T) \) as above. For brevity, put \( p = p_0 \) so that we write \( \lambda_1(p) \equiv \lambda_1(p, p; \{ T_{R,j} \}) \) as

\[
\lambda_1(p) = \sum_{j=1}^{J} \sum_{m \in \mathbb{N}_{R,j}} \tilde{T}(m_j, m_{-j}) \binom{R}{m_1, \cdots, m_J} m_1 p_1^{m_1} p_2^{m_2} \cdots p_{j-1}^{m_{j-1}} p_j^{m_j+1} p_{j+1}^{m_{j+1}} \cdots p_J^{m_J}.
\]

Let us compute \( c_1(m_1, \cdots, m_J) \). Then it suffices to show that \( c_j(m_1, \cdots, m_J) \) is the same for all \( j = 1, \cdots, J \), or, without loss of generality, that

\[
c_1(m_1, \cdots, m_J) = c_2(m_1, \cdots, m_J).
\]

First observe that \( c_2(m_1, m_2, \cdots, m_J) = c_1(m_2, m_1, \cdots, m_J) \) by the form of \( \tilde{T} \) in (18) and \( \Lambda_R \). Hence it suffices to show that

\[
c_1(m_1, m_2, \cdots, m_J) = c_1(m_2, m_1, \cdots, m_J).
\] (20)

To show this, first note that

\[
c_1(m_1, \cdots, m_J) = \tilde{T}(m_1, m_{-1}) \binom{R}{m_1, \cdots, m_J} m_1 + U_R
\] (21)

where

\[
U_R = \sum_{j=2}^{J} \tilde{T}(m_1 + 1, m_2, \cdots, m_{j-1}, m_j - 1, m_{j+1}, \cdots, m_J) \times \binom{R}{m_1 + 1, m_2, \cdots, m_{j-1}, m_j - 1, m_{j+1}, \cdots, m_J}(m_1 + 1).
\]

The relation in (21) holds for any \( (m_1, \cdots, m_J) \in \mathbb{N}_{R,J} \) and we can simply extend the domain of \( \tilde{T}(m_j, m_{-j}) \) to negative numbers by taking \( \tilde{T}(m_j, m_{-j}) = 0 \) if \( m_j < 0 \). By noting

\[
\binom{R}{m_1 + 1, m_2, \cdots, m_j - 1, m_{j+1}, \cdots, m_J}(m_1 + 1) = \binom{R}{m_1, \cdots, m_J} m_j,
\]

we write the coefficient \( c_1(m_1, \cdots, m_J) \) in (21) as

\[
\left( \binom{R}{m_1, \cdots, m_J} \right) \left[ m_1 \tilde{T}(m_1, m_{-1}) + \sum_{j \neq 1} m_j \tilde{T}(m_j - 1, m_1 + 1, m_2, \cdots, m_{j-1}, m_{j+1}, \cdots, m_J) \right]
\] (22)

and similarly, write \( c_1(m_2, m_1, m_3, \cdots, m_J) \) as

\[
\left( \binom{R}{m_1, \cdots, m_J} \right) \left[ m_2 \tilde{T}(m_2, m_{-2}) + \sum_{j \neq 2} m_j \tilde{T}(m_j - 1, m_1 + 1, m_2, \cdots, m_{j-1}, m_{j+1}, \cdots, m_J) \right]
\] (23)
Since the factor in front of the above brackets in (22) and (23) are the same, it suffices for (20) to show that

\[ m_1 \tilde{T}(m_1, m_{-1}) + \sum_{j \neq 1} m_j \tilde{T}(m_j - 1, m_1 + 1, m_2, \ldots, m_{j-1}, m_{j+1}, \ldots, m_j) = m_2 \tilde{T}(m_2, m_{-2}) + \sum_{j \neq 2} m_j \tilde{T}(m_j - 1, m_1 + 1, m_2, \ldots, m_{j-1}, m_{j+1}, \ldots, m_j). \]

By rearranging terms on both sides of the second equality, we obtain

\[ m_1 \left[ \tilde{T}(m_1; m_2, \ldots, m_J) - \tilde{T}(m_1 - 1; m_2 + 1, m_3, \ldots, m_J) \right] + \sum_{j=3}^{J} m_j \left[ \tilde{T}(m_j - 1; m_1 + 1, m_2, \ldots, m_J) - \tilde{T}(m_j - 1; m_2 + 1, m_1, \ldots, m_J) \right] = m_2 \left[ \tilde{T}(m_2; m_1, m_3, \ldots, m_J) - \tilde{T}(m_2 - 1; m_1 + 1, m_3, \ldots, m_J) \right]. \] (24)

Therefore, the proof is complete once we show that the above equality is satisfied by our choice of (33). One can check this equality immediately by considering each case: \( m_1 = m_2 = 0 \) and \( m_1, m_2 > 0 \) and \( m_1 > 0, m_2 = 0 \) and finally \( m_1 = 0, m_2 > 0 \). However, here we take a different route, showing how the form of (33) was discovered. In the proof we generate sufficient conditions for the equality in (24). Then these sufficient conditions lead to the solution of (33).

Without loss of generality, we assume \( m_1 \geq m_2 \) and \( m_3 \geq m_4 \geq \cdots \geq m_J \). If \( m_1 = m_2 = 0 \), the equality in (24) is trivially satisfied.

**Case 1)** \( m_1, m_2 > 0 \). Then, the condition (24) is satisfied if

\[ m_1 \left[ \tilde{T}(m_1; m_2, \ldots, m_J) - \tilde{T}(m_1 - 1; m_2 + 1, m_3, \ldots, m_J) \right] = 1, \] (25)

and

\[ \tilde{T}(m_j - 1; m_1 + 1, m_2, \ldots, m_J) - \tilde{T}(m_j - 1; m_2 + 1, m_1, \ldots, m_J) = 0. \] (26)

Restriction (26) implies that \( \tilde{T}(m_1; m_2, \ldots, m_J) \) depends on \( (m_2, \ldots, m_J) \) only through \( \nu(m_2, \ldots, m_J) \), the number of non-zero elements from the non-choices \( \{m_2, \ldots, m_J\} \). To see this, choose \( (m'_2, \ldots, m'_J) \) such that \( \nu(m'_2, \ldots, m'_J) = \nu(m_2, \ldots, m_J) \). Then, we can show that

\[ \tilde{T}(m_1; m_2, \ldots, m_J) = \tilde{T}(m_1; m'_2, \ldots, m'_J), \]

by repeating the process in (26) with adding and subtracting by 1 between two non-zero members from \( \{m_2, \ldots, m_J\} \).

Therefore we write

\[ \tilde{T}(m_1, m_2, \ldots, m_J) = \tilde{T}(m_1, \nu(m_2, \ldots, m_J)), \]

where \( \nu \) denotes the number of non-zero elements in the non-choice set. Using the observation in (26), (25) can be re-written as

\[ m_1 \left[ \tilde{T}(m_1; \nu(m_2, \ldots, m_J)) - \tilde{T}(m_1 - 1; \nu(m_2 + 1, m_3, \ldots, m_J)) \right] = 1, \] (27)

and note that \( \nu(m_2, \ldots, m_J) = \nu(m_2 + 1, m_3, \ldots, m_J) \). Hence we extract one condition for \( \tilde{T} \) that leads to
Therefore, we extract a condition for (31):
\[ \tilde{T}(m, \nu) - \tilde{T}(m - 1, \nu) = \frac{1}{m} \quad \text{for all possible } m. \] (28)

Case 2) \( m_1 > 0 \) and \( m_2 = 0 \). If further, \( m_3 = 0 \) then \( m_1 \) is simply \( R \). In this case,
\[
m_1 \left[ \tilde{T}(m_1; m_2, \cdots, m_J) - \tilde{T}(m_1 - 1; m_2 + 1, m_3, \cdots, m_J) \right]
= R \left[ \tilde{T}(R; m_2 = 0, \cdots, m_J = 0) - \tilde{T}(R - 1; m_2 + 1, m_3 = 0, \cdots, m_J = 0) \right] = 0
\] (29)
or
\[ \tilde{T}(R, 0) = \tilde{T}(R - 1, 1) \] (30)

If on the other hand \( m_3 > 0 \), we have from (12)
\[
m_1 \left[ \tilde{T}(m_1; m_2, \cdots, m_J) - \tilde{T}(m_1 - 1; m_2 + 1, m_3, \cdots, m_J) \right]
+ \sum_{j=3}^{J} m_j \left[ \tilde{T}(m_j - 1; m_1 + 1, m_2, \cdots, m_J) - \tilde{T}(m_j - 1; m_2 + 1, m_1, \cdots, m_J) \right]
= 0.
\]

By subtracting and adding back \( \tilde{T}(m_1 - 1; m_2, m_3 + 1, \cdots, m_J) \), we can write the above equation as
\[
m_1 \left[ \tilde{T}(m_1; m_2, m_3, \cdots, m_J) - \tilde{T}(m_1 - 1; m_2, m_3 + 1, \cdots, m_J) \right]
= m_1 \left[ \tilde{T}(m_1 - 1; m_2 + 1, m_3, \cdots, m_J) - \tilde{T}(m_1 - 1; m_2, m_3 + 1, \cdots, m_J) \right]
+ \sum_{j=3}^{J} m_j \left[ \tilde{T}(m_j - 1; m_2 + 1, m_1, \cdots, m_J) - \tilde{T}(m_j - 1; m_1 + 1, m_2, \cdots, m_J) \right]
\] (31)

Note that the left hand side in (31) is 1 by (27) and the difference in the number of non-zero elements in \( \tilde{T} \) for each difference term on the right-hand side is exactly 1. For example, \( \nu(m_2 + 1, m_3, \cdots, m_J) = \nu(m_2, m_3 + 1, \cdots, m_J) + 1 \) and \( \nu(m_1 + 1, m_2, \cdots, m_J) = \nu(m_1 + 1, m_2, \cdots, m_J) + 1 \). Therefore, if
\[ \tilde{T}(m, \nu) - \tilde{T}(m, \nu - 1) = c \]
for some \( c \) independent of \( m \) and \( \nu \), (31) is satisfied. In this case, (31) becomes
\[ 1 = \sum_{j=1}^{J} cm_j = cR \text{ or } c = \frac{1}{R}. \]

Therefore, we extract a condition for (31):
\[ \tilde{T}(m, \nu) - \tilde{T}(m, \nu - 1) = \frac{1}{R} \] (32)
for all \( m \) and \( \nu \). To summarize, conditions (28), (30), and (32) are sufficient for (24).

Now, if we define
\[ \tilde{T}(m_j, m_{-j}) = - \sum_{s=0}^{R-m_j-1} \frac{1}{R-s} + \frac{\nu(m_{-j})}{R}, \] (33)
this choice of $\tilde{T}$ satisfies conditions (28), (30), and (32), and hence the equation (24) follows, completing the
proof. On the other hand, it is also worth noting that the conditions (28), (30), and (32) for $\tilde{T}$ also lead to
the form of (33) up to an affine transform. This is the way the transform $T_R$ is determined.

**Proof of Lemma 2**: We first consider the case of $J = 3$. Recall

$$\Lambda_R(p, p_0; T_R) = \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,3}} \left( \frac{R}{m_1, m_2, m_3} \right) T_{R,j}(m_1, m_2, m_3) p_1^{m_1} p_2^{m_2} p_3^{m_3}$$

where $T_{R,j}(m_1, m_2, m_3) = \tilde{T}(m_j; m_{-j})$ with $\tilde{T}$ as defined in (33). Recall that $\lambda_j$ denotes the derivative of
$\Lambda_R(p, p_0; \{T_{R,j}\})$ with respect to $p_j$, so that

$$\lambda_1 - \lambda_3 = \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,3}} \left( \frac{R}{m_1 - 1, m_2, m_3 + 1} \right) T_{R,j}(m_1 - 1, m_2, m_3 + 1) (m_3 + 1) p_1^{m_1-1} p_2^{m_2} p_3^{m_3+1} \{m_1 > 0\}$$

By relabeling the terms ($m_3$ as $m_3 + 1$ and $m_1$ as $m_1 - 1$),

$$\lambda_3 = \sum_{j=1}^{J} p_{j0} \sum_{m \in \mathbb{N}_{R,3}} \left( \frac{R}{m_1 - 1, m_2, m_3 + 1} \right) T_{R,j}(m_1 - 1, m_2, m_3 + 1) (m_3 + 1) p_1^{m_1-1} p_2^{m_2} p_3^{m_3+1} \{m_1 > 0\}$$

Hence the difference $\lambda_1 - \lambda_3$ is equal to $\sum_{m \in \mathbb{N}_{R,3}} B_R(m) p_1^{m_1-1} p_2^{m_2} p_3^{m_3}$, where

$$B_R(m) = \sum_{j=1}^{3} p_{j0} \left( \frac{R}{m_1, m_2, m_3} \right) \{T_{R,j}(m_1, m_2, m_3) - T_{R,j}(m_1 - 1, m_2, m_3 + 1)\} m_1 \{m_1 > 0\}$$

However, by the definition of $T_{R,j}$, we have

$$m_1 \left[ T_{R,j}(m_1, m_2, m_3) - T_{R,j}(m_1 - 1, m_2, m_3 + 1) \right]$$

$$= m_1 \times 1\{j = 1\} \left[ \frac{1}{m_1} - \frac{1\{m_3 = 0\}}{R} \right] + m_1 \times 1\{j = 2\} \left[ \frac{1\{m_1 = 1\}}{R} - \frac{1\{m_3 = 0\}}{R} \right]$$

$$+ m_1 \times 1\{j = 3\} \left[ \frac{-1}{m_3 + 1} + \frac{1\{m_1 = 1\}}{R} \right].$$

Plugging this back into $B_R(m)$ we obtain

$$B_R(m) = p_{10} \left( \frac{R}{m_1, m_2, m_3} \right) [1 - 1\{m_3 = 0\} \frac{m_1}{R}]$$

$$+ p_{20} \left( \frac{R - 1}{m_1 - 1, m_2, m_3} \right) [1\{m_1 = 1\} - 1\{m_3 = 0\}]$$

$$+ p_{30} \left( \frac{R}{m_1 - 1, m_2, m_3 + 1} \right) [-1 + 1\{m_1 = 1\} \frac{m_3 + 1}{R}].$$
Now, write the summand in $\lambda_1 - \lambda_3$:

$$B_R(m)p_m^{m_1-1}p_2^{m_2}p_3^{m_3} = \left\{ \begin{array}{l} p_{10} \left( \frac{R}{m_1,m_2,m_3} \right) p_{10}^{m_1}p_2^{m_2}p_3^{m_3} - p_{10} \left( \frac{R-1}{m_1-1,m_2,0} \right) p_{10}^{m_1-1}p_2^{m_2} \right\} I(m_1 > 0) \\
+ p_{20} \left( \frac{R-1}{0,m_2,m_3} \right) p_{20}^{m_2}p_3^{m_3} - p_{20} \left( \frac{R-1}{m_1-1,m_2,0} \right) p_{20}^{m_1-1}p_2^{m_2}I(m_1 > 0) \\
- \frac{p_{30}}{p_3} \left( \frac{R}{m_1-1,m_2,m_3+1} \right) p_{30}^{m_1-1}p_2^{m_2}p_3^{m_3+1}I(m_1 > 0) + p_{30} \left( \frac{R-1}{0,m_2,m_3} \right) p_2^{m_2}p_3^{m_3}. \right\}$$

Summing the above over $m \in \mathbb{N}_{R,3}$ and rearranging the terms, we obtain that $\lambda_1 - \lambda_3$ is equal to

$$\frac{p_{10}}{p_1} \left[ 1 - \sum_{m \in \mathbb{N}_{R,3}} \left( \frac{R}{0,m_2,m_3} \right) p_{10}^{m_2}p_3^{m_3} \right] - p_{10} \left( \frac{p_1 + p_2}{R-1} \right)^R - p_{20} \left( \frac{p_2 + p_3}{R-1} \right)^R - p_{20} \left( \frac{p_1 + p_2}{R-1} \right)^R$$

or

$$\frac{p_{10}}{p_1} \left[ 1 - \sum_{m \in \mathbb{N}_{R,3}} \left( \frac{R}{m_1-1,m_2,0} \right) p_{10}^{m_1-1}p_2^{m_2} \right] + p_{30} \left( \frac{p_2 + p_3}{R-1} \right)^R - \left( \frac{p_1 + p_2}{R-1} \right)^R.$$

Using the fact that $p_1 + p_2 + p_3 = 1$ and $p_{10} + p_{20} + p_{30} = 1$, we find that the above becomes,

$$\frac{p_{10}}{p_1} \left[ 1 - (1 - p_1)^R \right] + (1 - p_{10}) (1 - p_1)^R$$

$$- \frac{p_{30}}{p_3} \left[ 1 - (1 - p_3)^R \right] - (1 - p_{30}) (1 - p_3)^R$$

$$= \frac{p_{10}}{p_1} \left( 1 - \frac{p_{10}}{p_1} \right) (1 - p_1)^R - \frac{p_{30}}{p_3} (1 - p_3)^R - \left( 1 - \frac{p_{30}}{p_3} \right) (1 - p_3)^R.$$

Therefore, $\partial (\lambda_1 - \lambda_3) / \partial p_1$ is equal to

$$\lambda_{11} - \lambda_{31} = \frac{p_{10}}{p_1} \left( 1 - \frac{p_{10}}{p_1} \right) (1 - p_1)^R - \left( 1 - \frac{p_{10}}{p_1} \right) (R - 1) (1 - p_1)^R - 2$$

and by symmetry, $\partial (\lambda_3 - \lambda_1) / \partial p_3$ is equal to

$$\lambda_{33} - \lambda_{31} = \frac{p_{30}}{p_3} \left( 1 - \frac{p_{30}}{p_3} \right) (1 - p_3)^R - \left( 1 - \frac{p_{30}}{p_3} \right) (R - 1) (1 - p_3)^R - 2.$$

We also obtain that $\partial (\lambda_1 - \lambda_3) / \partial p_2 = \lambda_{12} - \lambda_{32} = 0$. Likewise, from

$$\lambda_2 - \lambda_3 = \frac{p_{20}}{p_2} \left( 1 - \frac{p_{20}}{p_2} \right) (1 - p_2)^R - \frac{p_{30}}{p_3} (1 - p_3)^R - \left( 1 - \frac{p_{30}}{p_3} \right) (1 - p_3)^R - 1,$$
we obtain
\[
\lambda_{22} - \lambda_{32} = -\frac{p_{20}}{p_2^2} + \frac{p_{20}}{p_2^2} (1 - p_2)^{R-1} - \left(1 - \frac{p_{20}}{p_2}\right) (R - 1) (1 - p_2)^{R-2},
\]
\[
\lambda_{33} - \lambda_{32} = -\frac{p_{30}}{p_3^2} + \frac{p_{30}}{p_3^2} (1 - p_3)^{R-1} - \left(1 - \frac{p_{30}}{p_3}\right) (R - 1) (1 - p_3)^{R-2},
\]
\[
\lambda_{21} - \lambda_{31} = 0.
\]

Note also that \(\lambda_{13} - \lambda_{33} = \lambda_{23} - \lambda_{33}\). We write
\[
\Lambda_R(p, p_0; T_R) = \Lambda_R((p_1, p_2, p_3), p_0; T_R) = \Lambda_R((p_1, p_2, 1 - p_1 - p_2), p_0; T_R) = \tilde{\Lambda}_R(\tilde{p}, p_0; T_R).
\]

Viewing this as a function of \((p_1, p_2)\), we find that its Hessian matrix is given by
\[
H_3 = \begin{pmatrix}
\lambda_{11} - \lambda_{31} - (\lambda_{13} - \lambda_{33}) & \lambda_{21} - \lambda_{31} - (\lambda_{23} - \lambda_{33}) \\
\lambda_{12} - \lambda_{32} - (\lambda_{13} - \lambda_{33}) & \lambda_{22} - \lambda_{32} - (\lambda_{23} - \lambda_{33})
\end{pmatrix}
\]
\[
= \begin{pmatrix}
\lambda_{11} - \lambda_{31} & \lambda_{21} - \lambda_{31} \\
\lambda_{12} - \lambda_{32} & \lambda_{22} - \lambda_{32}
\end{pmatrix}
- \begin{pmatrix}
\lambda_{13} - \lambda_{33} & \lambda_{13} - \lambda_{33} \\
\lambda_{13} - \lambda_{33} & \lambda_{13} - \lambda_{33}
\end{pmatrix}
= \begin{pmatrix}
\lambda_{11} - \lambda_{31} & 0 \\
0 & \lambda_{22} - \lambda_{32}
\end{pmatrix}
- (\lambda_{13} - \lambda_{33}) \begin{pmatrix}
1 & 1 \\
1 & 1
\end{pmatrix}.
\]

Note that \(\lambda_{11} - \lambda_{31}\) is only a function of \(p_{10}\) and \(p_1\). We want to show
\[
\lambda_{11} - \lambda_{31} = -\frac{p_{10}}{p_1^2} + \frac{p_{10}}{p_1^2} (1 - p_1)^{R-1} - \left(1 - \frac{p_{10}}{p_1}\right) (R - 1) (1 - p_1)^{R-2}
\]
\[
= \frac{p_{10}}{p_1} h(R) - (R - 1)(1 - p_1)^{R-2},
\]
where \(h(R) = (1 - p_1)^{R-1} + p_1(R - 1)(1 - p_1)^{R-2}\). We rewrite
\[
h(R) = e^{(R-2)\log(1-p_1)} (1 + (R - 2)p_1) - 1.
\]

Now the first order derivative of \(h(R)\) is given by
\[
h'(R) = e^{(R-2)\log(1-p_1)} \{\log(1-p_1)(1 + (R - 2)p_1) + p_1\}
\]
\[
\leq e^{(R-2)\log(1-p_1)} \{\log(1-p_1) + p_1\} \leq 0
\]
when \(R \geq 2\), because \(\log(1-p_1) + p_1 \leq 0\) for \(p_1 \in [0, 1]\). This implies that
\[
\lambda_{11} - \lambda_{31} \leq \max \left\{ \frac{p_{10}}{p_1} h(2) - 1, \frac{p_{10}}{p_1} h(3) \right\} = \max \{-1, -p_{10}\} \leq -p_{10} \leq -\varepsilon.
\]
Similarly, $\lambda_{22} - \lambda_{32} \leq -\varepsilon$ and $\lambda_{33} - \lambda_{13} \leq -\varepsilon$. Therefore, from (34),

$$a^\top H_3 a \leq -2\varepsilon \left( \sum_{j=1}^{J-1} a_j^2 \right).$$

Consider the case $J > 3$. First, we get

$$\lambda_j - \lambda_k = \frac{p_{j0}}{p_j} + \left( 1 - \frac{p_{j0}}{p_j} \right) (1 - p_j)^{R-1} - \frac{p_{k0}}{p_k} - \left( 1 - \frac{p_{k0}}{p_k} \right) (1 - p_k)^{R-1},$$

for all $j, k = 1, 2, ..., J$. Then it suffices to check the negative definiteness of the matrix

$$\begin{pmatrix}
\lambda_{11} - \lambda_{J1} & 0 & \ldots & 0 \\
0 & \lambda_{22} - \lambda_{J2} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \lambda_{J-1,J-1} - \lambda_{J,J-1}
\end{pmatrix} - (\lambda_{1J} - \lambda_{JJ})
\begin{pmatrix}
1 & 1 & \ldots & 1 \\
1 & 1 & \ldots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \ldots & 1
\end{pmatrix}.$$

And as before, it suffices to show that $\lambda_{11} - \lambda_{J1} < 0$ for all $p_1$ and $p_{J0}$, because then, by symmetry, $\lambda_{jj} - \lambda_{J,j} < 0$ for all $j = 1, 2, ..., J - 1$. This can be proved exactly in the same way as before. ■

**Proof of Theorem 1**: Take $\theta \in \Theta$ such that $||\theta - \theta_0|| > \eta$, for some $\eta > 0$. Then

$$P\{|p_i(\theta) - p_i(\theta_0)|| > 0\} > 0.$$  

Now, we focus on $E[\Lambda_R(p_i(\theta), p_i(\theta_0); T_R)]$. We write $\tilde{p}_i(\theta) = (p_1(X_i; \theta), \ldots, p_{J-1}(X_i; \theta))^\top$ and $\lambda_j = \partial \Lambda_R(\tilde{p}, p_0; T_R)/\partial \tilde{p}_j$. Then

$$E[\Lambda_R(p_i(\theta), p_i(\theta_0); T_R)] = E[\Lambda_R(\tilde{p}_i(\theta), p_i(\theta_0); T_R)]$$

$$= \sum_{j=1}^{J-1} E\left[ \lambda_j(\tilde{p}_i(\theta_0), p_i(\theta_0); T_R)(\tilde{p}_j(X_i; \theta) - \tilde{p}_j(X_i; \theta)) \right]$$

$$+ E\left[ (\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0))^\top \left( \frac{\partial^2 \Lambda_R(p_i^*, p_i(\theta_0); T_R)}{\partial \tilde{p} \partial \tilde{p}} \right) (\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0)) \right],$$

where $p_i^*$ lies on the line segment connecting $\tilde{p}_i(\theta)$ and $\tilde{p}_i(\theta_0)$. The first term is zero because $\lambda_j(p_i(\theta_0), p_i(\theta_0); T_R) = \lambda_j(p_i(\theta_0), p_i(\theta_0); T_R)$ by Lemma 1 and $\sum_{j=1}^{J-1} p_j(X_i; \theta) = \sum_{j=1}^{J-1} p_j(X_i; \theta_0) = 1$. As for the second term, by Lemma 2, it is negative semidefinite. Hence we can bound it from below by

$$E\left[ (\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0))^\top \left( \frac{\partial^2 \Lambda_R(p_i^*, p_i(\theta_0); T_R)}{\partial \tilde{p} \partial \tilde{p}} \right) (\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0)) \right]$$

$$\geq \sum_{j=1}^{J-1} E\left[ (\tilde{p}_i(\theta) - \tilde{p}_i(\theta_0))^2 \right] \varepsilon.$$

The last term is positive because $P\{\tilde{p}_i(\theta) \neq \tilde{p}_i(\theta_0)\} > 0$. ■
The proofs of Theorems 2-3 below require the following preliminary results. Define for $\theta \in \Theta$,

$$T_{ij}^*(\theta) \equiv T_R(m_{ij}^*(\theta), m_{-ij}^*(\theta)) \quad \text{and} \quad \Delta_{ij}(\theta) \equiv T_{ij}^*(\theta) - T_{ij}(\theta_0).$$

**Lemma A1:** Suppose that Assumptions 1-2 hold. Then for $\theta \in \Theta$,

$$\sum_{j=1}^{J} \mathbb{E}(D_{ij}\Delta_{ij}(\theta)) = \sum_{j=1}^{J} (\theta - \theta_0)\Omega_j(\theta_\ast)^\top (\theta - \theta_0),$$

where $\theta_\ast$ lies on the line segment between $\theta$ and $\theta_0$ and $\Omega_j(\theta) = \frac{\partial^2}{\partial \theta \partial \theta} \mathbb{E}(D_{ij}\Delta_{ij}(\theta)))$.

**Proof:** By the mean-value theorem,

$$\sum_{j=1}^{J} \mathbb{E}(D_{ij}\Delta_{ij}(\theta)) = \sum_{j=1}^{J} \psi_j(\theta_0)^\top (\theta - \theta_0) + \sum_{j=1}^{J} (\theta - \theta_0)\Omega_j(\theta_\ast)^\top (\theta - \theta_0)$$

where $\theta_\ast$ lies on the line segment between $\theta$ and $\theta_0$ and $\psi_j(\theta) = \frac{\partial}{\partial \theta} \mathbb{E}(D_{ij}\Delta_{ij}(\theta))$. However,

$$\sum_{j=1}^{J} \psi_j(\theta_0) = \frac{\partial}{\partial \theta} \sum_{j=1}^{J} \mathbb{E}(D_{ij}\Delta_{ij}(\theta)|_{\theta=\theta_0}) = \frac{\partial}{\partial \theta} \sum_{j=1}^{J} \mathbb{E}(D_{ij}(T_j^*(\theta) - T_{ij}^*(\theta_0)))|_{\theta=\theta_0}$$

$$= \sum_{j=1}^{J} \frac{\partial}{\partial \theta} \mathbb{E}[\Lambda_R(p_{ij}(\theta), p_{ij}(\theta_0); T_R) - \Lambda_R(p_{ij}(\theta_0), p_{ij}(\theta_0); T_R)|_{\theta=\theta_0} = 0,$$

because $\mathbb{E}[\Lambda_R(p_{ij}(\theta), p_{ij}(\theta_0); T_R)]$ is maximized at $\theta = \theta_0$ by Theorem 1. Hence we obtain the wanted result.

**Lemma A2:** Suppose that Assumptions 1-2 hold. Then for $\delta > 0$,

$$\sup_{\theta \in B(\theta_0; M\delta)} \mathbb{E}l_{n,R}^*(\theta) - \mathbb{E}l_{n,R}^*(\theta_0) \leq C \log(R)\delta^2$$

for some $C > 0$.

**Proof:** First, observe that from Lemma A1, $\mathbb{E}l_{n,R}^*(\theta) - \mathbb{E}l_{n,R}^*(\theta_0)$ is less than or equal to

$$\frac{1}{2} \left[ \sup_{\theta \in B(\theta_0, \delta)} (\theta - \theta_0)^\top \frac{\partial^2}{\partial \theta \partial \theta} \mathbb{E} \left[ \sum_{j=1}^{J} D_{ij} T_{ij}^*(\theta) \right] (\theta - \theta_0) \right].$$

Note that

$$\frac{\partial^2}{\partial \theta \partial \theta} \mathbb{E}[D_{ij} T_{ij}^*(\theta)] = -\frac{1}{R} \sum_{m=0}^{R-1} \frac{\partial^2}{\partial \theta \partial \theta} \mathbb{E} \left[ D_{ij} 1\{1 - m/R > p_{ij}^*(\theta)\} \right] + \frac{1}{R} \frac{\partial^2}{\partial \theta \partial \theta} \mathbb{E}[\nu(m_{ij}^*(\theta))].$$

We consider the first term. We let $H_{R,T_j}$ be the CDF of the binomial distribution with parameter $(R, p_j)$.
Then define
\[ H^{(1)}_{R,p_j}(\cdot) = \frac{\partial H_{R,p_j}(\cdot)}{\partial p_j} \quad \text{and} \quad H^{(2)}_{R,p_j}(\cdot) = \frac{\partial^2 H_{R,p_j}(\cdot)}{\partial p_j^2} \]
We write the first term in (36) as
\[ -\frac{1}{R} \sum_{m=0}^{R-1} \frac{\partial^2}{\partial \theta \partial \theta^\top} E \left( D_{ij} 1\{R - m > \sum_{m=1}^{R} \delta_j(X_i, \eta_{i,m}; \theta)\} \right) \]
\[ = E \left[ -\frac{p_j(X_i, \theta_0)}{R} \sum_{m=0}^{R-1} H^{(1)}_{R,p_j(X_i, \theta)}(R - m - 1) \frac{\partial^2 p_j(X_i, \theta)}{\partial \theta \partial \theta^\top} \right] \]
\[ + E \left[ -\frac{p_j(X_i, \theta_0)}{R} \sum_{m=0}^{R-1} H^{(2)}_{R,p_j(X_i, \theta)}(R - m - 1) \frac{\partial p_j(X_i, \theta) \partial p_j(X_i, \theta)}{\partial \theta \partial \theta^\top} \right] \]
where \( H^{(1)}_{R,p_j(X_i, \theta)}(\cdot) \) and \( H^{(2)}_{R,p_j(X_i, \theta)}(\cdot) \) are the first order and the second order derivatives of the binomial distribution function with parameter \((R, p_j(X_i, \theta))\). By Assumption 1(ii), we have \( p(X_i, \theta) \in B(p(X_i, \theta_0), C(X_i)\delta) \) for all \( \theta \in B(\theta_0, \delta) \) where \( C(X_i) \) is square integrable and does not depend on \( \theta \in B(\theta_0, \delta) \). By taking \( \delta \) small, we have eventually \( 1 \geq p_j(X_i, \theta) > \varepsilon > 0 \) for the constant \( \varepsilon > 0 \) in Assumption 2(iii) with large probability. For this \( p(X_i, \theta) \), the derivatives \( H^{(1)}_{R,p_j(X_i, \theta)}(\cdot) \) and \( H^{(2)}_{R,p_j(X_i, \theta)}(\cdot) \) are bounded uniformly over \( \theta \in B(\theta_0, \delta) \) with large probability. Using Assumption 1(ii), we can bound the Euclidean norm of the first term in (36) by
\[ \sup_{\theta \in B(\theta_0, \delta)} \sqrt{E \left[ \left( \frac{\partial^2 p_j(X_i, \theta)}{\partial \theta \partial \theta^\top} \right)^2 \right]} \leq C \cdot \frac{1}{R} \sum_{m=0}^{R-1} \frac{1}{1-m/R} \]
And note that
\[ C \frac{1}{R} \sum_{m=0}^{R-1} \frac{1}{1-m/R} \leq C \int_0^{1-1/R} \frac{1}{1-u} du + O(R^{-1}) = \log(R) + O(R^{-1}). \]
Hence the first term on the right-hand side of (36) is \( O(\log(R)) \).
Now we consider the second term in (36). Note that
\[ \frac{\partial^2}{\partial \theta \partial \theta^\top} E[\nu(m_{i,j}(\theta))] = \sum_{k=1, k \neq j}^J \frac{\partial^2}{\partial \theta \partial \theta^\top} E \left[ P \left\{ \sum_{m=1}^{R} \delta_k(X_i, \eta_{i,m}; \theta) > 0 \right\} \right] \]
\[ = \sum_{k=1, k \neq j}^J \frac{\partial^2}{\partial \theta \partial \theta^\top} E \left[ \{ 1 - H_{R,p_k(X_i, \theta)}(0) \} \right] \]
\[ = \sum_{k=1, k \neq j}^J E \left[ -H^{(2)}_{R,p_k(X_i, \theta)}(0) \frac{\partial p_j(X_i, \theta) \partial p_j(X_i, \theta)}{\partial \theta \partial \theta^\top} - H^{(1)}_{R,p_k(X_i, \theta)}(0) \frac{\partial^2 p_j(X_i, \theta)}{\partial \theta \partial \theta^\top} \right]. \]
As argued before, we can take \( \delta \) small so that for all \( \theta \in B(\theta_0, \delta) \), \( 1 > p(X_i, \theta) > \varepsilon > 0 \) for some \( \varepsilon \). And this leads to the fact that \( H^{(s)}_{R,p_j(X_i, \theta)}(0) \) and \( H^{(s)}_{R,p_j(X_i, \theta)}(0) \), \( s = 1, 2 \), are bounded uniformly over \( \theta \in B(\theta_0, \delta) \). Therefore, the Euclidean norm of the second term in (36) is again bounded by
\[ \sup_{\theta \in B(\theta_0, \delta)} \sqrt{E \left[ \left( \frac{\partial^2 p_j(X_i, \theta)}{\partial \theta \partial \theta^\top} \right)^2 \right]} + C \times \frac{1}{R}. \]
Hence we conclude $\sup_{\theta \in B(\theta_0, \delta)} \left| \mathbf{E} l^*_{n,R}(\theta) - \mathbf{E} l^*_{n,R}(\theta_0) \right| \leq C \log(R) \delta^2$. We obtain (35). 

**Proof of Theorem 2:** We first show the consistency of the estimator. Given the identification result, it suffices for consistency to show that

$$\sup_{\theta \in \Theta} \left| l^*_{n,R}(\theta) - l_R(\theta) \right| \rightarrow_p 0 \text{ as } n \rightarrow \infty,$$

where $l_R(\theta) = \mathbf{E} l^*_{n,R}(\theta)$. Since $\Theta$ is compact and for each $\theta \in \Theta$ we have

$$l^*_{n,R}(\theta) \rightarrow_p l_R(\theta),$$

by the Law of Large Numbers, it suffices to show the following stochastic equicontinuity condition, i.e., for any $\varepsilon, \eta > 0$, there exists $\delta > 0$ such that for each $\tilde{\theta} \in \Theta$,

$$P \left\{ \sup_{\theta \in B(\tilde{\theta}, \delta)} \left| l^*_{n,R}(\theta) - l^*_{n,R}(\tilde{\theta}) \right| > \eta \right\} < \varepsilon.$$

Define $T_R$ as in (33). Recall that we can write $l^*_{n,R}(\theta)$ as

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_R(m_{ij}^*(\theta), m_{ij}^*(\theta)) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \left[ \sum_{m=0}^{R-1} \frac{1 \{ m \leq R - m_{ij}^*(X_i, \eta_i^*; \theta) \} - 1} {R - m} \right] + \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \nu(m_{ij}^*(\theta)) \nu^{-1}.$$

Hence, $\left| l^*_{n,R}(\theta) - l^*_{n,R}(\tilde{\theta}) \right|$ is bounded by $C \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \{ m_{ij}^*(X_i, \eta_i^*; \theta) - m_{ij}^*(X_i, \eta_i^*; \tilde{\theta}) \} \geq \eta \},$ for any small $\eta > 0$, because $m_{ij}^*(X_i, \eta_i^*; \theta)$ is an integer. (Note that $C$ may depend on $R$.) Therefore, for each $\tilde{\theta} \in \Theta$,

$$P \left\{ \sup_{\theta \in B(\tilde{\theta}, \delta)} \left| l^*_{n,R}(\theta) - l^*_{n,R}(\tilde{\theta}) \right| > \eta \right\} \leq J P \left\{ \sup_{\theta \in B(\tilde{\theta}, \delta)} \sup_{1 \leq j \leq J} \frac{1}{n} \sum_{i=1}^{n} \left| m_{ij}^*(X_i, \eta_i^*; \theta) - m_{ij}^*(X_i, \eta_i^*; \tilde{\theta}) \right| \geq \eta / (CJ) \right\}.$$

However, by Assumption 2(ii), $E \left[ \sup_{\theta \in B(\tilde{\theta}, \delta)} \left| m_{ij}^*(X, \eta; \theta) - m_{ij}^*(X, \eta; \tilde{\theta}) \right| \right]$ is bounded by

$$\sum_{r=1}^{R} E \left[ \sup_{\theta \in B(\tilde{\theta}, \delta)} \left| \delta_j(X_i, \eta_i^*; \tilde{\theta}) - \delta_j(X_i, \eta_i^*; \tilde{\theta}) \right| \right] \leq R \delta^{1/2}.$$

This yields the stochastic equicontinuity of the process $l^*_{n,R}(\tilde{\theta})$ and thereby, completes the proof for the consistency of $\tilde{\theta}$.

Now we turn to the rate of convergence. Since we have Lemma A2, in view of Theorem 3.2.5 of van der Vaart and Wellner (1996), it suffices for the completion of the proof to investigate the continuity modulus of the process $\sqrt{n} \{ l^*_{n,R}(\theta) - \mathbf{E} l^*_{n,R}(\theta) \}$. Given our definition of $T_R$, the objective function $l^*_{n,R}(\theta)$ can be
Proof of Theorem 3:

 rewritten as

\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} \delta_j(X_i, \eta_i; \theta_0) h_R(p_{ijR}(X_i, \theta), \nu(m_{-ij}^*(\theta))),
\]

where \( h_R(p, \nu) = -\frac{1}{R} \sum_{m=0}^{R-1} 1\{1 - m/R > p\} / (1 - (m/R)) + \nu/R. \) In the meanwhile,

\[
E \left[ \sup_{\theta : ||\theta - \theta_0|| \leq \delta} \left| h_R(p_{ijR}(X_i, \theta), \nu(m_{-ij}^*(\theta))) - h_R(p_{ijR}(X_i, \theta_0), \nu(m_{-ij}^*(\theta_0))) \right|^2 \right] \leq C \delta,
\]

by Assumption 2(ii). Let us define \( \gamma_j(D, X, \eta; \theta) = Dh_R(m_j(X, \eta; \theta) / R, \nu(m_j(X, \eta; \theta))) \) and \( G = \{ \gamma(\cdot, \cdot, \cdot, \cdot) : \theta \in \Theta \}. \) From the proof of Theorem 3.1 in Chen, Linton, and van Keilegom (2003), the result of (38) gives us

\[
\int_0^1 \sqrt{1 + \log N_\|2\| \|G\|^2, \Theta, \| \cdot \| \|^2} d\varepsilon \leq \int_0^1 \sqrt{1 + \log N_\|2\| \|G\|^2, \Theta, \| \cdot \| \|^2} d\varepsilon < \infty,
\]

where \( G \) is an envelope of \( G \). We define \( G_\delta = \{ \gamma_1 - \gamma_2 : \gamma_1, \gamma_2 \in G, \| \gamma_1 - \gamma_2 \|_2 < \delta \}. \) Then by the maximal inequality in terms of the bracketing entropy (e.g. Pollard (1989), van der Vaart (1996)), we have

\[
E \left[ \sup_{\theta \in B(\theta_0, \delta)} \sqrt{n} |l_{n,R}^*(\theta) - l_{n,R}^*(\theta_0) - \mathbb{E}l_{n,R}^*(\theta) + \mathbb{E}l_{n,R}^*(\theta_0)| \right] \leq C \int_0^1 \sqrt{1 + \log N_\|2\| \|G_\delta\|^2, \Theta, \| \cdot \| \|^2} d\varepsilon \|G_\delta\|^2 \leq C \|G_\delta\|^2,
\]

where \( G_\delta \) indicates the envelope of \( G \). The second inequality follows from the inequality:

\[
N_\|2\| \|G_\delta\|^2, \Theta, \| \cdot \| \|^2 \leq N_\|2\| \|G\|^2, \Theta, \| \cdot \| \|^2,
\]

and (39). By the result of (38), we can take \( G_\delta \) such that \( \|G_\delta\|^2 \leq C\delta^{1/2} \), and deduce that the continuity modulus of \( l_{n,R}^*(\theta) \) in \( \theta \) turns out to be \( O(\delta^{1/2}) \). Now, by Theorem 3.2.5 of van der Vaart and Wellner (1996), the rate of convergence \( r_n \) for \( \hat{\theta} \) satisfies \( r_n^{2 - 1/2} \leq \sqrt{n} \). Hence \( r_n \sim n^{1/3} \), yielding the result of the theorem.

**Proof of Theorem 3**: By Assumption 2(ii),

\[
P \left\{ \sup_{\theta \in B(\theta_0, \delta)} |p_{ij}^*(\theta) - p_{ij}(\theta)| > \epsilon/p \middle| X_i = x \right\} \leq \frac{2}{\epsilon p} E \left[ \sup_{\theta \in B(\theta_0, \delta)} |p_{ij}^*(\theta) - p_{ij}(\theta)| \middle| X_i = x \right]
\]

\[
\leq \frac{2}{\epsilon p \sqrt{R}} \int_0^{C} \sqrt{1 + \log N_\|2\| \|G\|^2, \Theta, \| \cdot \| \|^2} d\varepsilon = O(R^{-1/2}).
\]

The last inequality uses Assumption 2(ii) and the maximal inequality of Pollard (1989). Hence, for sufficiently
small $\delta > 0$,

$$\begin{align*}
P \left\{ \inf_{\theta \in B(\theta_0, \delta)} p_{ij}^*(\theta) > \frac{\epsilon_p}{2} \mid X_i = x \right\} \\
\geq P \left\{ \inf_{\theta \in B(\theta_0, \delta)} p_{ij}^*(\theta) > \frac{\epsilon_p}{2} + \sup_{\theta \in B(\theta_0, \delta)} \left| p_{ij}^*(\theta) - p_{ij}(\theta) \right| \mid X_i = x \right\} \\
\geq P \left\{ \frac{\epsilon_p}{2} > \sup_{\theta \in B(\theta_0, \delta)} \left| p_{ij}^*(\theta) - p_{ij}(\theta) \right| \mid X_i = x \right\} \to 1 \text{ for every } x,
\end{align*}$$

because $P \left\{ \inf_{\theta \in B(\theta_0, \delta)} p_{ij}(\theta) > \frac{\epsilon_p}{2} \mid X_i \right\} = 1$ by Assumption 2(iii). Hence $\inf_{\theta \in B(\theta_0, \delta)} p_{ij}^*(\theta) > \frac{\epsilon_p}{2}$ with conditional probability given $X_i$ converges to one almost surely. We assume that $\inf_{\theta \in B(\theta_0, \delta)} p_{ij}^*(\theta) > \frac{\epsilon_p}{2}$ for the rest of the proof.

**Claim 1:**

$$E \left[ \sup_{\theta \in B(\theta_0, \delta)} \sqrt{n} \left| l_{n,R}^*(\theta) - l_{n,R}(\theta_0) - E l_{n,R}^*(\theta) + E l_{n,R}(\theta_0) \right| \right] \leq \mu_R(\delta),$$

where

$$\mu_R(\delta) = C \left\{ \delta^{1/2} \sqrt{-\log \delta / \sqrt{R}} + \delta \right\} \times \left\{ \sqrt{\log(R)} + \sqrt{\log(-\log \delta)} \right\} + C \sqrt{R} R^{-1}.$$

Combining the results, we can establish the $\sqrt{n}$-rate of convergence as we demonstrate now. Suppose we have shown Claim 1. Take a sequence $r_n = n^{1/2}$ and partition $\Theta$ into "shells" $R_{j,n} = \{ \theta : 2^{j-1} < r_n | \theta - \theta_0 | \leq 2^j \}$ with $j$ ranging over integers. For any $\eta, M > 0$, we have

$$P \left\{ r_{n,\theta} > 2^M \right\} \leq \sum_{j \geq M} \sum_{2^j \leq \eta_{r_n}} P \left\{ \sup_{\theta \in R_{j,n}} \left| l_{n,R}^*(\theta) - l_{n,R}(\theta_0) \right| \geq \eta \right\} \leq P \left\{ 2 ||\hat{\theta} - \theta_0|| \geq \eta \right\}.$$  \hspace{1cm} (41)

The second probability on the right-hand side vanishes because $\hat{\theta}$ is consistent. For each $\theta \in R_{j,n}$, we have

$$E l_{n,R}^*(\theta) - E l_{n,R}(\theta_0) \leq \frac{C 2^{j-2} \log(R)}{r_n^2}$$

by Lemma A2. By using Claim 1, the sum of probabilities on the right-hand side in (41) is bounded by

$$\begin{align*}
\sum_{j \geq M} \sum_{2^j \leq \eta_{r_n}} P \left\{ \sup_{\theta \in R_{j,n}} \left| l_{n,R}^*(\theta) - l_{n,R}(\theta_0) + E l_{n,R}^*(\theta) \right| \geq \frac{C 2^{j-2} \log(R)}{r_n^2} \right\} \\
\leq C \sum_{j \geq M} \sum_{2^j \leq \eta_{r_n}} \frac{C \sqrt{n}}{\sqrt{R} \log(R)} \frac{2^{j-1}}{\sqrt{n}} \\
\leq C \frac{\sqrt{n}}{\sqrt{R} \log(R)} \sum_{j \geq M} \frac{C}{\sqrt{R} \log(R)} \sum_{2^j \leq \eta_{r_n}} 2^{-3j/2} + C \frac{\sqrt{n}}{\sqrt{R} \log(R)} \sum_{2^j \leq \eta_{r_n}} 2^{-j+1} + C \frac{\sqrt{n}}{\sqrt{R} \log(R)} \sum_{2^j \leq \eta_{r_n}} 2^{-2j+2} \to 0,
\end{align*}$$

as $M \to \infty$, because $n^{1/4} \sqrt{\log n} / \sqrt{R} \log(R) \leq C n^{1/4} \sqrt{\log R} / \sqrt{R} \log(R) = C n^{1/4} / \sqrt{R} < \infty$. Therefore, $\sqrt{n}$-consistency of the estimator $\hat{\theta}$ follows.

Having established the $\sqrt{n}$-consistency of $\hat{\theta}$, we establish the asymptotic normality of the estimator as follows. We first show the following.
Claim 2: There exists a sufficiently large $M > 0$ such that

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} [D_{ij} \Delta_{ij}(\theta) - \mathbb{E}(D_{ij} \Delta_{ij}(\theta))] \tag{42}
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} \left\{ D_{ij}(p_j(X_i, \theta) - p_j(X_i, \theta_0)) \right\} - \mathbb{E} \left[ \frac{D_{ij}(p_j(X_i, \theta) - p_j(X_i, \theta_0))}{p_j(X_i, \theta_0)} \right] + o_P(1),
\]

uniformly over $\theta \in B(\theta_0, n^{-1/2})$.

Suppose that we have shown Claim 2. Let $V(X_i) = \sum_{j=1}^{J} D_{ij} \frac{\partial}{\partial \theta_0} p_j(X_i, \theta_0)/p_j(X_i, \theta_0)$. From Lemma A1 and Claim 2, we write

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} [D_{ij} \Delta_{ij}(\theta) - \mathbb{E}(D_{ij} \Delta_{ij}(\theta))] = (\theta - \theta_0)^\top Z_n + o_P(n^{-1/2}), \tag{43}
\]

uniformly over $\theta \in B(\theta_0, Mn^{-1/2})$, as $n \to \infty$ and $R \to \infty$ jointly, where $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V(X_i)$. Note that $\mathbb{E} V(X_i) = 0$. Now, we follow similar steps in the proof of Theorem 3.2.16 in van der Vaart and Wellner (1996). Combined with Claim 2, the result of (43) yields

\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^*(\hat{\theta}) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^*(\theta_0) \tag{44}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} \Delta_{ij}(\hat{\theta}) - \frac{1}{2} (\hat{\theta} - \theta_0)^\top \Omega (\hat{\theta} - \theta_0) + \frac{1}{\sqrt{n}} (\hat{\theta} - \theta_0)^\top Z_n + o_P(n^{-1}).
\]

Similarly

\[
\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^*(\theta_0 - n^{-1/2} \Omega^{-1} Z_n) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} D_{ij} T_{ij}^*(\theta_0) \tag{45}
\]

\[
= - \frac{1}{2n} Z_n^\top \Omega^{-1} Z_n + o_P(n^{-1}).
\]

By the definition of $\hat{\theta}$, the left-hand side of (44) is larger than the left-hand side of (45). We subtract the second equation from the first equation to obtain

\[
\frac{1}{2} (\hat{\theta} - \theta_0 + n^{-1/2} \Omega^{-1} Z_n)^\top \Omega (\hat{\theta} - \theta_0 + n^{-1/2} \Omega^{-1} Z_n) \geq -o_P(n^{-1}).
\]

Since $\Omega$ is negative definite, we conclude that

\[
\sqrt{n} (\hat{\theta} - \theta_0) = \Omega^{-1} Z_n + o_P(1).
\]

The wanted result follows by the usual CLT. The proof of the theorem is complete.
Proof of Claim 1: First, observe that for \( p, p_0 \in (\varepsilon_p/2, 1) \),

\[
\frac{1}{R} \sum_{m=0}^{R-1} \frac{1\{1 - m/R > p\} - 1\{1 - m/R > p_0\}}{1 - (m/R)}
\]

(46)

\[
= \int_{1-p_0}^{1-p} \frac{1}{1-u} du + O \left( R^{-1} \right)
\]

\[
= \log(p) - \log(p_0) + O \left( R^{-1} \right) = \frac{p - p_0}{p_0} + O(|p - p_0|^2 + R^{-1}).
\]

Let \( \tilde{T}^*_ij(\theta) = \log(p^*_ij(\theta)) - \log(p^*_ij(\theta_0)) \). Since \( \inf_{\theta \in B(\theta_0, \delta)} p^*_ij(\theta) > \varepsilon_p/2 \), we deduce that

\[
E \left[ \sup_{\theta \in B(\theta_0, \delta)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} (D_{ij}T^*_ij(\theta) - D_{ij}\tilde{T}^*_ij(\theta_0) - E [D_{ij}T^*_ij(\theta) - D_{ij}\tilde{T}^*_ij(\theta_0)]) \right]
\]

\[
= E \left[ \sup_{\theta \in B(\theta_0, \delta)} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} (D_{ij}\tilde{T}^*_ij(\theta) - D_{ij}\tilde{T}^*_ij(\theta_0) - E [D_{ij}\tilde{T}^*_ij(\theta) - D_{ij}\tilde{T}^*_ij(\theta_0)]) \right]
\]

\[+ O(\sqrt{n}R^{-1} + \sqrt{n}\delta^2). \]

We focus on the last expectation. Observe that

\[
\sqrt{E \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| D_{ij}\tilde{T}^*_ij(\theta) - D_{ij}\tilde{T}^*_ij(\theta_0) \right|^2 \right]} \leq C \sqrt{E \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p^*_ij(\theta) - p^*_ij(\theta_0) \right|^2 \right]}
\]

\[\leq C \sqrt{E \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p^*_ij(\theta) - p^*_ij(\theta_0) - \{p^*_ij(\theta_0) - p^*_ij(\theta_0)\} \right|^2 \right]}, \]

\[+ C \sqrt{E \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p^*_ij(\theta) - p^*_ij(\theta_0) \right|^2 \right]}. \]

Using Theorem 2.14.5 of van der Vaart and Wellner (1996), the leading term is bounded by

\[
CE \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p^*_ij(\theta) - p^*_ij(\theta_0) - \{p^*_ij(\theta_0) - p^*_ij(\theta_0)\} \right| \right] + C\delta^{1/2}/\sqrt{R}.
\]

As for the leading expectation above, we proceed similarly as in (40): for all \( x \),

\[
E \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| p^*_ij(\theta) - p^*_ij(\theta_0) - \{p^*_ij(\theta_0) - p^*_ij(\theta_0)\} \right| \right] = C \delta^{1/2} \int_0^{C^{1/2}} \sqrt{1 + \log N(\|\varepsilon/C\|^2, \Theta, \|\cdot\|)} d\varepsilon \leq C\delta^{1/2} \sqrt{-\log \delta/\sqrt{R}},
\]

using Assumption 2(ii) and the maximal inequality. Therefore,

\[
\sqrt{E \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| D_{ij}\tilde{T}^*_ij(\theta) - D_{ij}\tilde{T}^*_ij(\theta_0) \right|^2 \right]} \leq C\delta^{1/2} \sqrt{-\log \delta/\sqrt{R}} + C\delta.
\]

This inequality reveals both a bound for an envelope for the class of functions indexing the empirical process in Claim 1 and the local uniform \( L_2 \)-continuity condition for this process. (e.g. Chen, Linton, and van
Keilegom (2003).) Using the maximal inequality and after some algebra,

\[ E \left[ \sup_{\theta \in B(\theta_0, \delta)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{J} (D_{ij} \hat{T}^*_ij(\theta) - D_{ij} T^*_ij(\theta_0) - E \left[ D_{ij} \hat{T}^*_ij(\theta) - D_{ij} T^*_ij(\theta_0) \right] \right| \right] \]

\[ \leq C \int_0^{C^\delta/2} \frac{\sqrt{-\log \delta/\sqrt{R} + C \delta}}{\sqrt{1 - C \log \{\sqrt{R}/\sqrt{-\log \delta} \}} \delta^2} \right] \right] \times \left\{ \sqrt{\log(R)} + \sqrt{\log(-\log \delta)} \right\}. \]

**Proof of Claim 2:** Let \( \delta_n = Mn^{-1/2} \) for some large \( M \). Define \( Y_i = (X_i, \eta_{i,1}, \cdots, \eta_{i,R}) \) and

\[ g_R(Y_i, \theta) = \frac{p_{ij}^*(\theta) - p_{ij}^*(\theta_0)}{p_{ij}^*(\theta_0)} - \frac{p_{ij}(\theta) - p_{ij}(\theta_0)}{p_{ij}(\theta_0)}. \]

Since we have (similarly as in the proof of Claim 1)

\[ \sup_{\theta \in B(\theta_0, \delta_n)} |p_{ij}^*(\theta) - p_{ij}^*(\theta_0)| = O_P(\delta_n^{1/2}R^{-1/2}\sqrt{-\log \delta_n + \delta_n}) = O_P(\delta_n \sqrt{-\log \delta_n}) \]

we write (using (46)),

\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [D_{ij} \Delta_{ij}(\theta) - E(D_{ij} \Delta_{ij}(\theta))] \]

\[ = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ D_{ij} \left[ \frac{p_{ij}(\theta) - p_{ij}(\theta_0)}{p_{ij}(\theta_0)} \right] - E \left( D_{ij} \left[ \frac{p_{ij}(\theta) - p_{ij}(\theta_0)}{p_{ij}(\theta_0)} \right] \right) \right\} + D_n(\theta) + O_P(\delta_n^2 (-\log \delta_n)) + O_P(\sqrt{n}R^{-1}), \]

where

\[ D_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [D_{ij} g_R(Y_i, \theta) - E(D_{ij} g_R(Y_i, \theta))]. \]

Now, once we show that \( D_n(\theta) = o_P(1) \) uniformly over \( \theta \in B(\theta_0, \delta_n) \), the proof is complete.

Define \( G_R(\delta_n) = \{ g_R(\cdot, \theta) : \theta \in B(\theta_0, \delta_n) \} \). Then note that

\[ E \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| g_R(Y_i, \theta) - g_R(Y_i, \theta_0) \right|^2 \right] \]

\[ = E \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \left( p_{ij}^*(\theta) - p_{ij}^*(\theta_0) \right) \frac{p_{ij}(\theta_0) - p_{ij}(\theta_0) - p_{ij}(\theta)}{p_{ij}(\theta_0)p_{ij}(\theta)} \right|^2 \right] \]

\[ \leq E \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \left( p_{ij}(\theta_0) - p_{ij}(\theta_0) \right) \frac{p_{ij}(\theta) - p_{ij}(\theta)}{p_{ij}(\theta_0)p_{ij}(\theta)} \right|^2 \right] \]

\[ + E \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \frac{p_{ij}(\theta) - p_{ij}(\theta_0) - p_{ij}(\theta) - p_{ij}(\theta_0)}{p_{ij}(\theta_0)p_{ij}(\theta)} \right|^2 \right]. \]
Since we have \( \sup_{\theta \in B(\theta_0, \delta_n)} \| p_{ij}^*(\theta) - p_{ij}(\theta) \|^2 = O_P(R^{-1}) \), the first term is bounded by \( CR^{-1} \delta_n^2 \). Define \( d_{i,r}^*(\theta) = \delta(X_i, \eta_{i,r}^*, \theta) - \delta(X_i, \eta_{i,r}^*, \theta_0) \). Then by Theorem 2.14.5 in van der Vaart and Wellner (1996),

\[
\left( \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \frac{1}{R} \sum_{r=1}^{R} d_{i,r}^*(\theta) - \mathbb{E}[d_{i,r}^*(\theta)|X_i] \right|^2 |X_i| \right] \right)^{1/2} = CE \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \frac{1}{R} \sum_{r=1}^{R} d_{i,r}^*(\theta) - \mathbb{E}[d_{i,r}^*(\theta)|X_i] \right| |X_i| \right] + O(R^{-1/2} \delta_n^{1/2}) = O(R^{-1/2} \delta_n^{1/2} \sqrt{-\log \delta_n}).
\]

Hence \( \left( \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} |g_{R}(Y_i, \theta) - g_{R}(Y_i, \theta_0)|^2 \right] \right)^{1/2} \leq CR^{-1/2} \delta_n^{1/2} \sqrt{-\log \delta_n}. \) From this it follows that

\[
N(\varepsilon, G_R(\delta_n), || \cdot ||_2) \leq N[\varepsilon R^{1/2}/\sqrt{-\log \delta_n}^2, \Theta, || \cdot ||_2] \leq C(\varepsilon R^{-1/2} \sqrt{-\log \delta_n})^2.
\]

Now, we write

\[
\mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} |D_n(\theta)| \right] = \mathbb{E} \left[ \sup_{\theta \in B(\theta_0, \delta_n)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [D_{ij}g_R(Y_i, \theta) - \mathbb{E}(D_{ij}g_R(Y_i, \theta))] \right| \right] \\
\leq \int_0^{CR^{-1/2} \delta_n^{1/2} \sqrt{-\log \delta_n}} \sqrt{1 + \log N(\varepsilon, G_R(\delta_n), || \cdot ||_2)} d\varepsilon \to 0,
\]

because \( R^{-1/2} \delta_n^{1/2} \sqrt{-\log \delta_n} \leq R^{-1/2} \delta_n^{1/2} \sqrt{-\log(R)} \to 0. \) Therefore, \( \sup_{\theta \in B(\theta_0, \delta_n)} |D_n(\theta)| = o_P(1). \) We conclude that

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [D_{ij} \Delta_{ij}(\theta) - \mathbb{E}(D_{ij} \Delta_{ij}(\theta))] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ D_{ij} \left[ \frac{p_{ij}(\theta) - p_{ij}(\theta_0)}{p_{ij}(\theta)} \right] - \mathbb{E} \left( D_{ij} \left[ \frac{p_{ij}(\theta) - p_{ij}(\theta_0)}{p_{ij}(\theta_0)} \right] \right) \right\} + o_P(1).
\]

We obtain the wanted result.

\[
\text{References}
\]


