Sequential Estimation of Structural Models with Fixed Point Constraint

Hiroyuki Kasahara
Department of Economics
University of Western Ontario
hkasahar@uwo.ca

Katsumi Shimotsu
Department of Economics
Queen's University
shimotsu@econ.queensu.ca

April 10, 2008

Abstract

This paper considers the estimation problem of structural models of which empirical restrictions are characterized in terms of fixed point constraint, such as a structural dynamic discrete choice model and a model of dynamic games. We analyze the conditions under which the nested pseudo-likelihood (NPL) algorithm achieves convergence and derive its convergence rate. We find that the NPL algorithm may not necessarily converge when the fixed point mapping does not have a local contraction property. To address the issue of non-convergence, we propose alternative sequential estimation procedures that can achieve convergence even when the NPL algorithm does not and, upon convergence, some of our proposed estimation algorithms produce more efficient estimators than the NPL estimator. We also show that the similar convergence results hold for models with (time-varying) unobserved heterogeneity, where the EM algorithm is incorporated into the NPL algorithm. Furthermore, we extend the idea behind the NPL algorithm to the moment estimators, developing a recursive extension of popular two-step moment methods called the sequential generalized method of moments (GMM) algorithm. The sequential GMM algorithm has the convergence properties similar to those of the NPL algorithm.

Keywords: approximate maximum likelihood, contraction, dynamic games, nested generalized method of moments, nested pseudo likelihood, unobserved heterogeneity.

JEL Classification Numbers: C13, C14, C63.

1 Introduction

Empirical implications of economic theory are often characterized by fixed point problems. Upon estimating such models, researchers typically consider a class of extremum estimators with fixed point constraint:

$$\max_{\theta\in\Theta} \quad Q_n(P) \quad s.t. \quad P = \Psi(P,\theta), \tag{1}$$

where $Q_n(P) = n^{-1}\sum_{i=1}^{n} \ln P(Z_i)$ for maximum likelihood estimator (MLE, hereafter) while $Q_n(P) = -\left[n^{-1}\sum_{i=1}^{n} g(Z_i,P)\right]' \hat{W} \left[n^{-1}\sum_{i=1}^{n} g(Z_i,P)\right]$ for the generalized method of moments estimator (GMM, hereafter) with the moment condition $E[g(Z_i,P^0)] = 0$ evaluated at the true probability P^0. Here, $\{Z_i\}_{i=1}^{n}$ is the sample data drawn from P^0.

The fixed point constraint $P = \Psi(P,\theta)$ in (1) summarizes the set of structural restrictions of the model that is parametrized with a finite vector $\theta \in \Theta$. When the model is correctly specified, the probability distribution obtained as the fixed point of the operator Ψ evaluated at the true parameter θ^0 generates the sample data. The examples of operator $\Psi(\cdot,\theta)$ include the policy iteration operator for a single agent dynamic programming model (e.g., Rust (1987), Hotz and Miller (1993)), and an operator defined by best response function for games (e.g., Bajari, Benkard, and Levin (2007), Pakes, Ostrovsky and Berry (2007), Pesendorfer and Schmidt-Dengler (2007)).

In principle, we may estimate the parameter θ in (1) by repeatedly solving the fixed point P_θ of $P = \Psi(P,\theta)$ at each parameter value to maximize the objective function with respect to θ. The major practical obstacle of applying such an estimation procedure lies in the computational burden because solving the fixed point problem for a given parameter can be very costly.

To reduce the computational burden, Hotz and Miller (1993) developed a simpler two-step estimator that does not require solving the fixed point problem for each trial value of the parameters. A number of recent papers in empirical industrial organization build on the idea of Hotz and Miller (1993) to develop two-step estimators for models with multiple agents (cf., Bajari, Benkard, and Levin, 2007; Pakes, Ostrovsky, and Berry, 2007; Pesendorfer and Schmidt-Dengler, 2007; Bajari, Chernozhukov, and Hong, 2006). These two-step estimators may suffer from substantial finite sample bias, however, when the choice probabilities are poorly estimated in the first step. This drawback is especially severe in estimating models with unobserved heterogeneity because it is difficult to obtain consistent initial estimates of choice probabilities.

To address the limitations of two-step estimators, Aguirregabiria and Mira (2002)(2007, AM07 hereafter) develop a recursive extension of the two-step method of Hotz and Miller (1993), called the *nested pseudo likelihood (NPL) algorithm* as follows. Starting from an initial estimate $\tilde{P}_0$, their algorithm iterates

Step 1: Given $\tilde{P}_{j-1}$, update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^{n} \ln[\Psi(\tilde{P}_{j-1},\theta)](Z_i)$.

Step 2: Update $\tilde{P}_{j-1}$ using the obtained estimate $\tilde{\theta}_j$: $\tilde{P}_j = \Psi(\tilde{P}_{j-1},\tilde{\theta}_j)$.

until $j = k$. AM07 show that their method can be applied to models with unobserved heterogeneity in the context of dynamic discrete games, and the NPL estimator—defined as the limit of the sequence generated by the NPL algorithm—is more efficient than the two-step estimators *if the convergence is achieved.*

While AM07 have obtained convergence in their simulations and illustrate that the NPL estimator performs very well relative to the two-step estimator, they do not provide the conditions under which the NPL algorithm converges. On the other hand, the simulation results of Pesendorfer and Schmidt-Dengler (2007) provide some evidence that the NPL algorithm may not necessarily converge while Collard-Wexler (2006) finds that $\tilde{P}_j$'s "cycle around several values without converging" in the NPL algorithm. If the NPL algorithm produces a sequence of estimators that diverge away from the fixed point of $P = \Psi(P, \theta)$ under the true parameter value θ^0, the algorithm may not be used to estimate the parameter in practice. To date, the convergence property of the NPL algorithm is not known in the literature.

This paper analyzes the conditions under which the NPL algorithm achieves convergence. The key to understanding the convergence properties of the NPL algorithm is a *contraction* property of the operator Ψ defining the fixed point problem. Intuitively, the faster the operator achieves contraction, the closer the the value obtained after one iteration is to the fixed point, and, therefore, we expect that the NPL algorithm works well if the operator has good contraction property. We show that the convergence of the NPL algorithm is achieved if the dominant eigenvalue of the Jacobian matrix $\partial\Psi(P, \theta)/\partial P$ evaluated at the fixed point P_θ is less than one in absolute value.[1] This is because the local contraction property of the operator Ψ is determined by the eigenvalues of the derivative of Ψ with respect to P. The closer the dominant eigenvalue of $\partial\Psi(P_\theta, \theta)/\partial P$ to zero, the faster the convergence rate of the NPL algorithm.

The violation of the condition that guarantees the convergence of the NPL algorithm is a concern. Using the dynamic discrete game model of AM07, our simulation results show that, when the degree of strategic substitutabilities is sufficiently high in dynamic game, the Jacobian matrix of the policy iteration mapping could have the smallest eigenvalue that is less than -1, leading to no convergence of the NPL algorithm. In such cases, various two step estimators can be used but they may suffer from the finite sample bias and are difficult to apply to models with unobserved heterogeneity.

We propose alternative sequential estimators that are implementable even when the original NPL algorithm does not converge. First, we consider modifying the fixed point mapping Ψ so that its transformed mapping shares the same fixed point as Ψ but has better contraction property. Upon convergence, the NPL algorithm with the transformed mapping produces an estimator that is characterized by the same first order conditions as the original NPL estimator.

Second, to further improve convergence property as well as efficiency, we propose a new estimation algorithm in which a pseudo-likelihood objective function in the NPL algorithm is

[1] In the context of single agent dynamic programming model, Kasahara and Shimotsu (2006, KS06 hereafter) derive the rate at which the sequence of the estimators generated from the NPL algorithm approaches the MLE. We extend the results of KS06 to a general class of structural models that are formulated as fixed point problem, including a model of dynamic games.

defined in terms of multiple iterations of the mapping as opposed to one iteration. In general, such a modification leads to a significant increase in the computational burden because repeated evaluations of the mapping is required for solving the optimization problem in Step 1. For this reason, we introduce an approximation method that requires evaluating the mapping and its Jacobian with respect to the parameter only once outside of the optimization routine. This algorithm converges faster than the original NPL algorithm and, when the convergence is achieved, the proposed estimator is more efficient than the NPL estimator.

Third, we also propose a sequential algorithm that may be used to obtain the maximum likelihood estimator upon convergence. The algorithm is based on directly approximating the fixed point of the mapping and is computationally attractive when an analytical expression for $\partial\Psi(P,\theta)/\partial P$ is available. Given an initially consistent estimator, taking one step of this sequential algorithm leads to an estimator that is asymptotically equivalent to the MLE. Taking additional steps produces a sequence of estimators that approaches the MLE in higher orders.

We also analyze the convergence properties of the NPL algorithm when it is applied to models with serially correlated unobserved state variables. Recently, Arcidiacono and Miller (2008) adapt the Expectation-Maximization (EM) algorithm into conditional choice probability estimator in the context of unobserved heterogeneity. When the conditional choice probabilities are updated between iterations of E-step and M-step, this algorithm may not be monotone increasing and the conditions for the algorithm to converge are not known. We show that the similar convergence results hold for models with time-varying unobserved heterogeneity, where the EM algorithm is incorporated into the NPL algorithm.

Finally, we develop a recursive extension of two-step *moment* estimators that are often used in estimating dynamic games, called the *sequential generalized method of moments (GMM) algorithm*. The sequential GMM algorithm replaces the pseudo-likelihood function in the NPL algorithm with the pseudo GMM objective function. We show that the convergence of the sequential GMM algorithm also requires that all the eigenvalues of $\partial\Psi(P_\theta,\theta)/\partial P$ are less than one in absolute value. The limit of the sequential GMM estimators may be more efficient than two-step estimators.

The reminder of the paper is organized as follows. Section 2 introduces a class of models with fixed point constraints. Section 3 establishes the convergence property of the NPL algorithm. In Section 4, we develop alternative sequential algorithms that can be used even when the NPL algorithm has convergence problems and, yet, achieve better asymptotic properties. Section 5 extends our analysis to the sequential GMM estimator. Section 6 applies our proposed methods to models with unobserved heterogeneity. Section 7 reports some simulation results.

2 The models with fixed point constraint and maximum likelihood estimator

We consider a class of parametric models of which restrictions are characterized in terms of fixed point problems in probability space. Upon estimating such models, researchers may consider the (conditional) maximum likelihood estimator (MLE) with fixed point constraint:

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \left\{ \max_{P\in\mathcal{M}_\theta} \; n^{-1} \sum_{i=1}^{n} \ln P(a_i|x_i) \right\}, \tag{2}$$

where

$$\mathcal{M}_\theta = \{P \in B_P : \; P = \Psi(P,\theta)\} \tag{3}$$

is a set of fixed points of $\Psi(\cdot,\theta)$ given the value of $\theta \in \Theta \subset \mathbb{R}^K$. Here, B_P represents the space of conditional probabilities while Θ is the set of possible parameter values. The model space—the set of probabilities that are consistent with the parametric fixed point restrictions—is then defined as a union of $\mathcal{M}_\theta$ over Θ: $\mathcal{M} = \cup_{\theta\in\Theta}\mathcal{M}_\theta = \{P \in B_P : \;\; P = \Psi(P,\theta), \;\theta \in \Theta\}$. We assume that the model is correctly specified so that the conditional probability in population, denoted by P^0, belongs to the model space $\mathcal{M}$, i.e., $P^0 \in \mathcal{M}$.

The fixed point constraint $P = \Psi(P,\theta)$ in (3) summarizes the set of structural restrictions of the model that is parametrized with a finite K dimensional vector θ. For each θ, an operator $\Psi(\cdot,\theta)$ maps the space of conditional choice probabilities into itself. When the model is correctly specified, the true probability distribution P^0 is the fixed point of the operator Ψ evaluated at the true parameter θ^0, from which the sample data is generated. The examples of operator $\Psi(\cdot,\theta)$ include the policy iteration operator for single agent dynamic programming models (e.g., Rust (1987), Hotz and Miller (1993), Aguirregabiria and Mira (2002)) and an operator defined by best response functions for dynamic games (e.g., Aguirregabiria and Mira (2007), Bajari, Benkard and Levin (2007), Pakes, Ostrovsky and Berry (2007), Pesendorfer and Schmidt-Dengler (2007)).

Example 1 (A dynamic discrete choice model and the policy iteration mapping) *An agent maximizes the expected discounted sum of utilities, $E[\sum_{j=0}^{\infty} \beta^j \{u(x_{t+j}, a_{t+j};\theta) + \epsilon_{t+j}(a_{t+j})\}|a_t, x_t;\theta]$, where x_t is an observable state variable and $\epsilon_t(a_t)$ is a state variable that are known to the agent but not to the researcher. The Bellman equation for this dynamic optimization problem is*

$$V(x) = \int \max_{a\in A} \left\{ u(x,a;\theta) + \epsilon(a) + \beta \sum_{x'\in X} V(x') f(x'|x,a;\theta) \right\} g(d\epsilon|x), \tag{4}$$

where $\beta \in (0,1)$ is the discount factor, $g(\epsilon|x)$ is the joint distribution of $\epsilon = \{\epsilon(j) : j \in A\}$ and $f(x'|x,a;\theta)$ is transition function. For each value of θ, we may compute the fixed point of the

Bellman equation and the conditional choice probability is given by

$$P_\theta(a|x) = \int 1\left\{a = \arg\max_{j\in A}\left[u(x,j;\theta) + \epsilon(j) + \beta\sum_{x'\in X} V_\theta(x')f(x'|x,j;\theta)\right]\right\} g(d\epsilon|x), \quad (5)$$

where V_θ is the fixed point of (4). Using the Hotz and Miller (1993)'s invertibility proposition, we may derive the policy iteration mapping in the space of conditional choice probability, $\Psi(\cdot,\theta)$, of which fixed point is the same as the conditional choice probabilities in (5), i.e., $P_\theta = \Psi(P_\theta,\theta)$ (cf., Aguirregabiria and Mira (2002) and Kasahara and Shimotsu (2006)).

Example 2 (A dynamic discrete choice model with finite dependence) *In a dynamic discrete choice model, the policy iteration mapping is not necessarily the only mapping to characterize fixed point constraint. Arcidiacono and Miller (2008) shows that, when the dynamic discrete choice problem exhibits finite time dependence, it is possible to derive an alternative mapping that is much simpler to compute than the policy iteration mapping. We illustrate their method by considering a simple machine replacement model of Rust (1987).*

Suppose $a \in \{0,1\}$ is the replacement decision for a bus engine, where $a = 1$ corresponds to replacing a bus engine. Let x denote the engine's mileage with $X = \{1,2,\ldots\}$. The transition function of x is given by $f(x_{t+1}|x_t,a_t)$ which takes a value of one for $x_{t+1} = (1-a_t)(x_t+1)+a_t$, and zero otherwise. In this case, the choice $a_t = 1$ is a renewal action and the model exhibits finite dependence. In particular, denoting

$$v(x,a) = u(x,a;\theta) + \beta\sum_{x'\in X} V(x')f(x'|x,a) \quad (6)$$

in (4), we have $v(x,1) = u(x,1;\theta) + \beta V(1)$. Then, assuming that $\epsilon(a)$ is independently drawn from the Type-I extreme-value distribution, (4) is written as $V(x) = v(x,1) - \ln P(a=1|x) + \gamma + \beta V(1)$, where γ is Euler's constant.[2] *Substituting this expression into the right hand side of (6) and taking a difference between $v(x,1)$ and $v(x,0)$ give*

$$v(x,1)-v(x,0) = [u(x,1;\theta)-u(x,0;\theta)]+\beta\sum_{x'\in X}[u(x',1;\theta)-\ln P(a=1|x')][f(x'|x,1)-f(x'|x,0)].$$

For each value of θ, the right hand side of this equation can be viewed as a mapping from the probability space to the space of value differences. Denote this mapping $\varphi(P,\theta)$. Then, we may derive an alternative mapping to the policy iteration mapping as

$$P(a=1|x) = \frac{\exp(v(x,1))}{\exp(v(x,1))+\exp(v(x,0))} = \frac{1}{1+\exp\left(-[\varphi(P,\theta)](x)\right)} \equiv [\Psi(P,\theta)](a=1|x). \quad (7)$$

[2] This follows from $V(x) = \gamma + \ln[\sum_{j=0}^{1}\exp(v(x,j))] = \gamma + u(x,1;\theta) - \ln P(a=1|x) = u(x,1) - \ln P(a=1|x) + \gamma + \beta V(1)$, where the second equality uses $P(a=1|x) = \exp(v(x,1))/[\sum_{j=0,1}\exp(v(x,j))]$.

Evaluating the mapping defined by the last term of (7) is much less computationally intensive than evaluating the policy iteration mapping, especially when the state space is large.

Example 3 (A dynamic discrete game) *Consider the model of dynamic discrete games studied by Aguirregabiria and Mira (2007). There are N global firms who are the potential entrants in M separate markets. At the beginning of each period, a firm makes an entry/exit choice in each market, i.e., $a_{it} \in A = \{0, 1\}$. The profit of firm i operating in period t depends on the vector of current firms' current decision $a_t = (a_{1t}, ..., a_{Nt})'$, the market demand condition S_t, its previous entry decision $a_{i,t-1}$, and the vector of firms's state that is private information to each firm $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_{Nt})'$. Let $\tilde{\Pi}_i(a_t, S_t, a_{t-1}, \epsilon_t; \theta)$ be firm i's profit in period t. Then, firm i maximizes the expected discounted sum of profits $E\left[\sum_{t=0}^{\infty} \beta^t \Pi_i(a_t, S_{mt}, a_{i,t-1}, \epsilon_{it}; \theta) | S_{mt}, a_{m,t-1}; \theta\right]$. We assume that S_t follows an exogenous first-order Markov process $f_S(S_{t+1}|S_t, a_{t-1}; \theta)$, which is common knowledge while ϵ_{it} is iid across markets and firms conditional on S_t and a_{t-1}.*

Let $\sigma^*(\theta) = \{\sigma_i^*(S_t, a_{t-1}, \epsilon_{it}; \theta) : i = 1, \ldots, N\}$ denote a set of strategy functions in a stationary Markov perfect equilibrium (MPE) given θ. Then, the equilibrium conditional choice probabilities are given by

$$P_i^{\sigma^*(\theta)}(a_i|S_t, a_{t-1}) = \int 1\{a_i = \sigma_i^*(S_t, a_{t-1}, \epsilon; \theta)\} g(d\epsilon|S_t, a_{t-1}), \tag{8}$$

where $g(\epsilon|S_t, a_{t-1})$ is the conditional distribution function for $\epsilon = \{\epsilon(a) : a \in A\}$. Aguirregabiria and Mira (2007) provides a best response mapping in probability space of which fixed point is identical to the equilibrium conditional choice probabilities in (8) so that $P_\theta = \Psi(P_\theta, \theta)$ where $P_\theta = \{P_i^{\sigma^*(\theta)} : i = 1, ..., N\}$.

The computation of the maximum likelihood estimator (MLE) in (2) requires repeatedly solving all the fixed points of $P = \Psi(P, \theta)$ at each parameter value to maximize the objective function with respect to θ. When there are multiple fixed points, finding all the fixed points of $P = \Psi(P, \theta)$ may be computationally infeasible. Even if there is a unique fixed point for each θ, the MLE could be extremely computationally intensive when evaluating the mapping Ψ is costly. For example, the MLE is often impractical in estimating models of dynamic game in example 3 with the modest number of players since the state space increases at exponential rate as the number of players increases. One of the major econometric issues in estimating models with fixed point constraint is to develop an estimator that is computational simple and has good finite sample properties as an alternative to the MLE.

3 The nested pseudo likelihood algorithm

3.1 Asymptotic properties of the NPL estimator

This section reviews the properties of the two-step pseudo maximum likelihood estimator (PML) and the estimator generated by the nested pseudo likelihood (NPL) algorithm as discussed in Aguirregabiria and Mira (2002, 2007). They are feasible alternatives to the MLE.

The pseudo maximum likelihood (PML) estimator is

$$\hat{\theta}_{PML} = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^{n} \ln \Psi(\hat{P}_0,\theta)(a_i|x_i),$$

where $\hat{P}_0$ is an initial consistent estimator for P^0.

We assume that the support of (a_i, x_i) is finite, $A \times X = \{a^1, a^2, ..., a^{|A|}\} \times \{x^1, x^2, \ldots, x^{|X|}\}$. Accordingly, P is represented with a $L \times 1$ vector while, given θ, the Jacobian $(\partial/\partial P')\Psi(P,\theta)$ is a $L \times L$ matrix, where $L = |A||X|$.

Assumption 1 *(a) Θ is compact and, for any $\theta \in \Theta$, $\mathcal{M}_\theta$ is compact. (b) $\Psi(P,\theta)$ is three times continuously differentiable. (c) $\Psi(P,\theta)(a|x) > 0$ for any (a,x) and any $\{P,\theta\} \in B_P \times \Theta$. (d) (a_i,x_i) for $i = 1,2,\ldots,N$, are independently and identically distributed, and $dF(x) > 0$ for any x in the support of x_i, where $F(x)$ is the distribution function of x_i. (e) There is a unique $\theta^0 \in int(\Theta)$ and a unique $P_{\theta^0} \in \mathcal{M}_{\theta^0}$ such that, for any $(a,x) \in A \times X$, $P_{\theta^0}(a|x) = P^0(a|x)$. For any $\theta \neq \theta^0$, $\Pr_{P^0}(\{(a,x) : \Psi(P^0,\theta)(a|x) \neq P^0(a|x)\}) > 0$. (g) $E_{\theta^0} \sup_{(P,\theta)} ||D^s\Psi(P,\theta)(a|x)||^2 < \infty$ for $s = 1,\ldots,4$.*

As shown in Proposition 1 of AM07, under Assumption 1, the two-step PML estimator is consistent and, when a root-n consistent estimator of P^0 is available, it is asymptotically normal.

Proposition 1 *Assume Assumption 1 holds and $\hat{P}_0 \to_p P^0$. Then $\hat{\theta}_{PML} \to_p \theta^0$.*

Proposition 2 *Assume Assumption 1 holds and $\sqrt{n}(\hat{P}_0 - P^0) \to_d N(0,\Sigma)$. Then, $\sqrt{n}(\hat{\theta}_{PML} - \theta^0) \to N(0, V_{PML})$, where $V_{PML} = (\Omega_{\theta\theta})^{-1} + (\Omega_{\theta\theta})^{-1}\Omega_{\theta P}\Sigma(\Omega_{\theta P})'(\Omega_{\theta\theta})^{-1}$ with*

$$\begin{aligned}\Omega_{\theta\theta} &\equiv E[(\partial/\partial\theta)\ln\Psi(P^0,\theta^0)(a|x)(\partial/\partial\theta')\ln\Psi(P^0,\theta^0)(a|x)] = -E[(\partial^2/\partial\theta\partial\theta')\ln\Psi(P^0,\theta^0)(a|x)],\\ \Omega_{\theta P} &\equiv E[(\partial/\partial\theta)\ln\Psi(P^0,\theta^0)(a|x)(\partial/\partial P')\ln\Psi(P^0,\theta^0)(a|x)] = -E[(\partial^2/\partial\theta\partial P')\ln\Psi(P^0,\theta^0)(a|x)].\end{aligned}$$

The second term of the variance expression, $(\Omega_{\theta\theta})^{-1}\Omega_{\theta P}\Sigma(\Omega_{\theta P})'(\Omega_{\theta\theta})^{-1}$, captures the effect of the first step estimator $\hat{P}_0$ on $\hat{\theta}_{PML}$. When the estimator $\hat{P}_0$ is imprecise as is often the case in practice, the two-step PML estimator may perform poorly. The eigenvalues of the Jacobian matrix $\Psi_P \equiv (\partial/\partial P')\Psi(P^0,\theta^0)$ is another important determinant of the variance V_{PML}. If all the eigenvalues of Ψ_P are equal to zero, then $\Omega_{\theta P} = 0$ and there is no effect of $\hat{P}_0$ on $\hat{\theta}_{PL}$ in the first order asymptotic. In this case, the limiting distribution of the two-step estimator is the same as that of the MLE (cf., Aguirregabiria and Mira (2002)), which is true even under the weaker assumption that $\hat{P}_0 - P^0 = O_p(n^{-b})$ with $b > 1/4$ (see Kasahara and Shimotsu (2006)).

Aguirregabiria and Mira (2002, 2007) consider a recursive extension of the two-step PML estimator based on the nested pseudo likelihood (NPL) algorithm as follows. Assume that an initial consistent estimator $\tilde{P}_0$ is available.

Step 1: Given $\tilde{P}_{j-1}$, update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^{n} \ln \Psi(\tilde{P}_{j-1}, \theta)(a_i|x_i)$.

Step 2: Update P using the obtained estimate $\tilde{\theta}_j$ by $\tilde{P}_j = \Psi(\tilde{P}_{j-1}, \tilde{\theta}_j)$.

Iterate Steps 1-2 until $j = k$.

This procedure generates a sequence of estimators $\{\tilde{P}_j, \tilde{\theta}_j\}_{j=1}^{k}$. If this sequence converges, its limit $(\hat{P}_{NPL}, \hat{\theta}_{NPL})$ is called the *NPL estimator*, satisfying the following two conditions:

$$\hat{\theta}_{NPL} = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^{n} \ln \Psi(\hat{P}_{NPL}, \theta)(a_i|x_i) \quad \text{and} \quad \hat{\theta}_{NPL} = \Psi(\hat{P}_{NPL}, \hat{\theta}_{NPL}). \tag{9}$$

The following proposition is from AM07 and states that $\hat{\theta}_{NPL}$ is root-n consistent asymptotically and more efficient than a two-step estimator if all the eigenvalues of Ψ_P are between 0 and 1.

Proposition 3 *Assume Assumption 1 holds. Then,* $\sqrt{n}(\hat{\theta}_{NPL} - \theta^0) \to N(0, V_{NPL})$, *where* $V_{NPL} = [\Omega_{\theta\theta} + \Omega_{\theta P}(I - \Psi_P)^{-1}\Psi_\theta]^{-1}\Omega_{\theta\theta}\{[\Omega_{\theta\theta} + \Omega_{\theta P}(I - \Psi_P)^{-1}\Psi_\theta]^{-1}\}'$ *with* $\Psi_\theta \equiv (\partial/\partial\theta')\Psi(P^0, \theta^0)$. *Furthermore, if all the eigenvalues of* Ψ_P *are less than one in absolute value, then* $V_{PML} - V_{NPL}$ *is positive definite.*

The estimator $\hat{\theta}_{NPL}$ can be obtained as a limit of iterating steps 1 and 2 *if the iterations converge.* Although AM07 have obtained convergence in their simulations and illustrate that the estimator $\hat{\theta}_{NPL}$ performs very well relative to the PML estimator, they neither provide the conditions under which the NPL algorithm converges nor analyze how fast the convergence occurs. On the other hand, some other studies find potential problems on the convergence of the NPL algorithm. The simulation results of Pesendorfer and Schmidt-Dengler (2007) provide some evidence that the NPL algorithm may not necessarily converge. Collard-Wexler (2006) uses the NPL method to estimate a structural model of entry and exit for the ready-mix concrete industry and finds that the NPL algorithm generates a sequence of $\hat{P}_j$'s that is oscillating without converging. To date, little is known about the convergence properties of the NPL algorithm.

3.2 Convergence properties of the NPL algorithm

We now analyze the conditions under which the NPL algorithm achieves convergence and derives its convergence rates. We show that its convergence property crucially depends on the eigenvalues of Ψ_P. In particular, if all the eigenvalues of Ψ_P are smaller than 1 in absolute value, then the NPL algorithm converges.

First, we state the regularity conditions. Denote $\overline{\psi}_\theta(P,\theta) = n^{-1}\sum_{i=1}^n(\partial/\partial\theta)\ln\Psi(P,\theta)(a_i|x_i)$, $\overline{\psi}_{\theta\theta}(P,\theta) = n^{-1}\sum_{i=1}^n(\partial^2/\partial\theta\partial\theta')\ln\Psi(P,\theta)(a_i|x_i)$, and $\overline{\psi}_{\theta P}(P,\theta) = n^{-1}\sum_{i=1}^n(\partial^2/\partial\theta\partial P')\ln\Psi(P,\theta)(a_i|x_i)$.

Assumption 2 *Assumption 1 holds, and in addition*

$$\begin{aligned} \overline{\psi}_\theta(P^0,\theta^0) &= O_p(n^{-1/2}), & \overline{\psi}_{\theta P}(P^0,\theta^0) &= -\Omega_{\theta P} + O_p(n^{-1/2}), \\ \overline{\psi}_{\theta\theta}(P^0,\theta^0) &= -\Omega_{\theta\theta} + O_p(n^{-1/2}), & \overline{\psi}_{\theta\theta}(P,\theta) &\textit{ is invertible for all } (P,\theta). \\ E\sup_{\theta,P}||D_{\theta P}\ln\Psi(P,\theta)|| &< \infty, & E\sup_{\theta,P}||D^3\ln\Psi(P,\theta)|| &< \infty, \\ \sup_{\theta,P}||D^2\Psi(P,\theta)|| &= O(1), \end{aligned}$$

All the assumptions but the last two are fairly weak. $\overline{\psi}_{\theta\theta}(P,\theta)$ should be invertible in many cases because $\overline{\psi}_{\theta\theta}(P,\theta)$ is an average of n matrices. If we assume $\tilde{P}_0$ is consistent, then the last assumption can be replaced by the invertibility of $\Omega_{\theta\theta}$.

Define $f_x(x_l) = \Pr(x = x^l)$ and let f_x be a $L \times 1$ vector of $\Pr(x = x^l)$ whose elements are arranged conformably with $P_{\theta^0}(a^j|x^l)$. Let $\Delta_P = diag(P^0)^{-1}diag(f_x)$. With these notations, we may write $\Omega_{\theta\theta} = \Psi_\theta'\Delta_P\Psi_\theta$ and $\Omega_{\theta P} = \Psi_\theta'\Delta_P\Psi_P$.

The following lemma is one of the main results of this paper. It states the local convergence rate of the NPL algorithm.

Lemma 1 *Suppose Assumption 2 holds. Then, for $j = 1,\ldots,k$,*

$$\begin{aligned} \tilde{\theta}_j - \hat{\theta}_{NPL} &= O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||), \\ \tilde{P}_j - \hat{P}_{NPL} &= M_{\Psi_\theta}\Psi_P(\tilde{P}_{j-1} - \hat{P}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2), \end{aligned}$$

where

$$M_{\Psi_\theta} \equiv I - \Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P.$$

It follows from induction that

$$\tilde{P}_k - \hat{P}_{NPL} = (M_{\Psi_\theta}\Psi_P)^k(\tilde{P}_0 - \hat{P}_{NPL}) + O((M_{\Psi_\theta}\Psi_P)^{k-1})[O_p(n^{-1/2}||\tilde{P}_0 - \hat{P}_{NPL}||) + O_p(||\tilde{P}_0 - \hat{P}_{NPL}||^2)].$$

If all the eigenvalues of $M_{\Psi_\theta}\Psi_P$ are less than 1 in absolute value, an iteration moves $\tilde{P}_j$ toward $\hat{P}_{NPL}$. Since the eigenvalues of M_{Ψ_θ} are either zero or one, the convergence property of $\tilde{P}_j$ is primarily determined by the dominant eigenvalues of Ψ_P.[3] That is, if all the eigenvalues

[3] In particular, we may show that, for any $y = P_1 - P_2$ where $P_1, P_2 \in B_P$,

$$||M_{\Psi_\theta}\Psi_P y|| \leq |\lambda({M_{\Psi_\theta}}'M_{\Psi_\theta})\lambda({\Psi_P}'\Psi_P)|^{1/2}||y||,$$

where $\lambda(Z)$ is the dominant eigenvalue of matrix Z.

of Ψ_P is sufficiently smaller than 1 in absolute value, then $\tilde{P}_k, \tilde{\theta}_k$ converges to $\hat{P}_{NPL}, \hat{\theta}_{NPL}$ as $k \to \infty$. In contrast, if some eigenvalues of $M_{\Psi_\theta}\Psi_P$ are larger than 1, then an iteration moves some elements of $\tilde{P}_j$ further away from $\hat{P}_{NPL}$. In this case, it is not clear whether the iterations eventually converge even when the initial estimate $\tilde{P}_0$ is in the neighborhood of $\hat{P}_{NPL}$.

Remark 1 $\Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P$ *is a generalized least squares projection matrix from a regression of an element of* B_P *onto the space spanned by* Ψ_θ*, where the "error variance matrix" is* Δ_P^{-1}*. On the other hand,* M_{Ψ_θ} *is the orthogonal projection matrix that generates the "residuals".*

Remark 2 *Even if the initial estimate,* $\tilde{P}_0$*, is not root-n consistent, iterations reduce the effect of the initial estimate on* $\tilde{\theta}_j$*, provided all the eigenvalues of* $M_{\Psi_\theta}\Psi_P$ *are smaller than 1 in absolute value.*

Remark 3 *If all the eigenvalues of* $M_{\Psi_\theta}\Psi_P$ *are smaller than 1 in absolute value and we choose* $k \to \infty$ *so that* $\log n = o(k)$*, then* $\tilde{P}_k - \hat{P}_{NPL} = o_p(n^{-1/2})$ *and the effect of* $\tilde{P}_0$ *on* $\hat{P}_{NPL}$ *vanishes in the limit. This is useful when some elements of* x *are continuously distributed and root-n consistent* $\tilde{P}_0$ *is not available.*

Remark 4 *When* $\Psi_P = 0$*, the convergence rate is faster than linear:*

$$\tilde{P}_j - \hat{P}_{NPL} = O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2).$$

Remark 5 *If at least one element of* x_i *is continuously distributed, one can prove the higher-order improvement by bootstrap as in Kasahara and Shimotsu (2006).*

4 Alternative sequential likelihood-based estimators

When a mapping $\Psi(P,\theta)$ is not a contraction in the neighborhood of (P^0,θ^0), the NPL algorithm has a convergence problem and therefore may not be used in practice. While the PML or other two-step estimators can be used in such cases, the finite sample bias is often a serious concern in these estimators. This section discusses alternative sequential algorithms that are implementable even when the NPL algorithm encounters a convergence problem. Some of our proposed estimators have better asymptotic properties than the NPL estimator.

4.1 Locally contractive mapping

In the previous section, we show that the dominant eigenvalue of Ψ_P is the main determinant of the convergence properties of the NPL algorithm. Expanding $\Psi(P,\theta^0)$ around the fixed point, $P^0 = P_{\theta^0}$, gives

$$\Psi(P,\theta^0) - P^0 = \Psi_P(P - P^0) + O(||P - P^0||^2)$$

so that the dominant eigenvalue of Ψ_P determines the rate of contraction for the mapping $\Psi(\cdot,\theta)$ in the neighborhood of (P^0,θ^0). If the dominant eigenvalue is less than 1 in absolute value, $\Psi(\cdot,\theta^0)$ is locally a contraction while, if it is more than 1, iterating $\Psi(\cdot,\theta^0)$ generates a sequence that does not converge to P^0. Thus, the convergence property of the NPL algorithm is determined by the local contraction property of $\Psi(\cdot,\theta)$ in the neighborhood of (P^0,θ^0).

In this section, we propose implementing the NPL algorithm by modifying the mapping $\Psi(P,\theta)$ so that its transformed mapping has better contraction property. We consider a class of mappings that are obtained as a log-linear combination of $\Psi(P,\theta)$ and P:

$$[\Lambda(P,\theta)](a|x) \equiv \{[\Psi(P,\theta)](a|x)\}^{\alpha} P(a|x)^{1-\alpha} \tag{10}$$

for all $(a,x) \in A \times X$, where $\alpha \in [0,1]$. Given θ, $\Lambda(P,\theta)$ is a mapping from B_P into itself. Since P is a fixed point of $\Psi(P,\theta)$ if and only if it is a fixed point of $\Lambda(P,\theta)$, we may obtain the fixed point of $\Psi(P,\theta)$ by solving the fixed point of $\Lambda(P,\theta)$. Furthermore, with an appropriate choice of α, the mapping $\Lambda(P,\theta)$ may become locally contractive with its dominant eigenvalue less than 1 even when the mapping $\Psi(P,\theta)$ is not locally contractive.

The following proposition states that, under certain conditions, we may choose the value of α so that the absolute value of the dominant eigenvalue of $\Lambda_P \equiv \nabla_{P'}\Lambda(P^0,\theta^0)$ is less than that of Ψ_P.

Proposition 4 *Denote the largest and the smallest eigenvalues of Ψ_P by $\lambda_{\max}$ and λ_{min}. If $\lambda_{\max} > 1 > \lambda_{\min}$, then there is no value of α such that all the eigenvalues of Λ_P are between -1 and 1. If $1 > \lambda_{\max} > \lambda_{\min}$, then the absolute value of the dominant eigenvalue of Λ_P is minimized at $\alpha^* = \frac{2}{2-\lambda_{\max}-\lambda_{\min}}$ and the largest and smallest eigenvalues of Λ_P are $\frac{\lambda_{\max}-\lambda_{\min}}{2-\lambda_{\max}-\lambda_{\min}}$ and $-\frac{\lambda_{\max}-\lambda_{\min}}{2-\lambda_{\max}-\lambda_{\min}}$, respectively, both of which are between -1 and 1. Furthermore, $\frac{\lambda_{\max}-\lambda_{\min}}{2-\lambda_{\max}-\lambda_{\min}}$ is smaller than the absolute value of the dominant eigenvalue of Ψ_P.*

We may consider the NPL algorithm using $\Lambda(P,\theta)$ in place of $\Psi(P,\theta)$ which iterates

Step 1: Update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^{n} \ln \Lambda(\tilde{P}_{j-1},\theta)(a_i|x_i)$ and

Step 2: Update P by $\tilde{P}_j = \Lambda(\tilde{P}_{j-1},\tilde{\theta}_j)$

until $j = k$. When the condition that $1 > \lambda_{\max} > \lambda_{\min}$ is satisfied, the sequence of estimators generated by the NPL algorithm with $\Lambda(P,\theta)$ may converge even if the NPL algorithm with $\Psi(P,\theta)$ does not converge. Furthermore, the limit of a sequence of estimators generated by the NPL algorithm with $\Lambda(P,\theta)$ satisfies the same first order conditions as that of (9) and it is identical to the original NPL estimator with $\Psi(P,\theta)$ upon convergence (see the Appendix B).

The advantage of this method is its simplicity. Once an appropriate value of α is determined, it achieves better convergence property than the original NPL algorithm without adding computational burden. The condition $1 > \lambda_{\max} > \lambda_{\min}$ may be restrictive in some cases but, in our Monte Carlo experiments using the model of Example 3, we find that $\lambda_{\max}$ is less than 1 while $\lambda_{\min}$ may become less than -1 when the degree of strategic substitutabilities is high.[4]

4.2 The q-NPL algorithm

Even when the absolute value of dominant eigenvalue of Λ_P or Ψ_P is strictly smaller than 1, the convergence of the NPL algorithm could be very slow and a sequence generated by the algorithm could behave erratically if the dominant eigenvalue is close to 1 in absolute value.[5]

In this section, we consider a possible extension of the NPL algorithm by defining a q-stage operator of Λ by

$$\Lambda^q(P,\theta) = \underbrace{\Lambda(\Lambda(...(\Lambda}_{\text{q times}}(P,\theta),\theta),...,\theta),\theta),$$

and $\Psi^q(P,\theta)$ is defined similarly. We define the q-NPL algorithm as the NPL algorithm using a q-stage operator Λ^q or Ψ^q in place of Λ or Ψ. In the following, we focus on the algorithm based on Λ^q but the same argument applies to Ψ^q. We recommend using Λ^q over Ψ^q in practice whenever the optimal value of α can be estimated because it improves convergence properties.

Given an initial consistent estimator $\tilde{P}_0$, the q-NPL algorithm iterates

Step 1: Update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \ln \Lambda^q(\tilde{P}_{j-1},\theta)(a_i|x_i)$ and

Step 2: Update P using the obtained estimate $\tilde{\theta}_j$ by $\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1},\tilde{\theta}_j)$

until $j = k$.

The limit of this sequence of estimators, denoted by $(\hat{P}_{qNPL},\hat{\theta}_{qNPL})$, satisfies

$$\hat{\theta}_{qNPL} = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^{n} \ln \Lambda^q(\hat{P}_{qNPL},\theta)(a_i|x_i) \quad \text{and} \quad \hat{\theta}_{qNPL} = \Lambda^q(\hat{P}_{qNPL},\hat{\theta}_{qNPL}), \tag{11}$$

if iterations converge. We call this estimator upon convergence the *q-NPL estimator*. Since the result of Lemma 1 also applies here by replacing Ψ with Λ^q, the dominant eigenvalue of $\Lambda_P^q \equiv (\partial/\partial P')\Lambda^q(P^0,\theta^0)$ is the main determinant of the convergence rate of the q-NPL algorithm. When the dominant eigenvalue of Λ_P, denoted by λ^*, is less than 1 in absolute value, the q-NPL algorithm converges faster than the NPL algorithm because the absolute value of dominant

[4] We may estimate the optimal choice, $\alpha^* = \frac{2}{2-\lambda_{\max}-\lambda_{\min}}$, by first applying the PML estimator and then evaluating the eigenvalues of $(\partial/\partial P')\Psi(\hat{P}_0,\hat{\theta}_{PML})$, which is a consistent estimator for Ψ_P. If it is difficult to evaluate the eigenvalues of $(\partial/\partial P')\Psi(\hat{P}_0,\hat{\theta}_{PML})$, we may simulate a sequence $\{P^j\}_{j=0}^J$ by iterating $P^j = \Psi(P^{j-1},\hat{\theta}_{PML})$ and compute the mean of $||P^{j+1}-P^J||/||P^j-P^J||$ across $j=1,...,J-1$, which gives an estimate of the dominant eigenvalue. Repeating this procedure for different values of α, say for $\alpha\in\{0.1,0.2,...,0.9\}$, we may estimate α^* by picking up the value of α that leads to the smallest value of the mean of $||P^{j+1}-P^J||/||P^j-P^J||$'s. We find that this procedure works well in our Monte Carlo experiments.

[5] As AM07 (pp.20-21) discuss, if some eigenvalues of Λ_P or Ψ_P are equal to one, then there could exist a continuum of NPL fixed points at (θ^0,P^0).

eigenvalue of Λ_P^q is equal to $|\lambda^*|^q$. Furthermore, the variance of the q-NPL estimator approaches to that of the MLE at the exponential rate of $|\lambda^*|^{2q}$ as $q \to \infty$. See the Appendix B.

A simple application of the q-NPL algorithm may not be so useful in practice, however, because computing Step 1 of the q-NPL algorithm requires repeatedly evaluating the mapping Λ at many different values of the vector of probabilities P. In contrast, an iteration of the NPL algorithm often requires evaluating the mapping Λ only once as discussed in Aguirregabiria and Mira (2002, 2007). For this reason, we consider the following *approximate q-NPL algorithm.*

Suppose that a consistent estimate $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ is available. Expanding $\Lambda^q(\tilde{P}_{j-1}, \theta)$ in Step 1 of the q-NPL algorithm gives

$$\Lambda^q(\tilde{P}_{j-1}, \theta) = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta'}\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\theta - \tilde{\theta}_{j-1}) + O(||\theta - \tilde{\theta}_{j-1}||^2). \tag{12}$$

Thus, $\Lambda^q(\tilde{P}_{j-1}, \theta)$ can be approximated by $\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta'}\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\theta - \tilde{\theta}_{j-1})$, and this approximation becomes exact as $\theta \to \tilde{\theta}_{j-1}$.

We propose to estimate θ using this approximation of $\Lambda^q(P, \theta)$. Let $(\tilde{P}_0, \tilde{\theta}_0)$ be an initial consistent estimator of (P^0, θ^0). For instance, $\tilde{\theta}_0$ can be the PML estimator. For $j \geq 1$, consider the following approximate q-NPL algorithm.

Step 1: Given $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, update θ by

$$\tilde{\theta}_j = \underset{\theta \in \Theta_j^q}{\arg\max}\, n^{-1} \sum_{i=1}^{n} \ln \tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_i|x_i),$$

where $\tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \equiv \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta'}\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\theta - \tilde{\theta}_{j-1})$ and $\Theta_j^q = \{\theta \in \Theta : \tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a|x) \in [\varepsilon, 1-\varepsilon]$ for all $(a, x) \in A \times X\}$ for an arbitrary small $\epsilon > 0$. We impose this restriction in order to avoid computing ln(0).[6]

Step 2: Given $(\tilde{\theta}_j, \tilde{P}_{j-1})$, update P using the obtained estimate $\tilde{\theta}_j$ by $\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_j)$.

Iterate Steps 1-2 until $j = k$.

Implementing Step 1 requires evaluating $\Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ and $\nabla_{\theta'}\Lambda^q(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ only once outside of the optimization routine for θ and, thus, it involves much fewer number of evaluations of $\Lambda(P, \theta)$ across different values of θ and P than the original q-NPL algorithm.[7] It still requires more iterations than the NPL algorithm but, when the NPL algorithm encounters convergence problems, this approximate q-NPL algorithm is feasible alternative.

[6] In practice, we may consider a penalized pseudo likelihood objective function by adding a penalty term that is increasing in the distance between $[\Lambda^q(\tilde{P}_0, \tilde{\theta}_0) + \nabla_{\theta'}\Lambda^q(\tilde{P}_0, \tilde{\theta}_0)(\theta - \tilde{\theta}_0)](a_i|x_i)$ and the set $[\varepsilon, 1-\varepsilon]$.

[7] $\Lambda^q(\tilde{P}_0, \tilde{\theta}_0)$ can be computed by just iterating $\Lambda(\tilde{P}_0, \tilde{\theta}_0)$ q times while $\nabla_{\theta'}\Lambda^q(\tilde{P}_0, \tilde{\theta}_0)$ can be computed by taking a numerical derivative of $\Lambda^q(\tilde{P}_0, \tilde{\theta}_0)$ with respect to the parameter vector θ. Using one-sided numerical derivatives, Step 1 requires $(K+1)q$ policy iterations.

To establish the consistency of the sequence of estimators generated by the approximate q-NPL algorithm, we need the following assumption in addition to Assumption 1.

Assumption 3 *For any $\eta \in \mathbb{R}^K$ such that $\eta \neq 0$, $\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a|x)\eta \neq 0$ with positive probability.*

Assumption 3 is an identification condition for the probability limit of our objective function and is required because we use an approximation of $\Lambda^q(P,\theta)(a|x)$ in the objective function. If this assumption is violated, then there exists a direction of θ such that $\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a|x)(\theta-\theta^0) = 0$ even when $\theta \neq \theta^0$. Then, it is not possible to identify θ^0. Assumption 3 is satisfied if the following $|X| \times K$ matrix has full column rank:

$$\begin{bmatrix} \nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a|x = X_1) \\ \vdots \\ \nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a|x = X_{|X|}) \end{bmatrix}.$$

Since $|X| \gg K$ in general, this condition is likely to be satisfied in most cases.

Under these assumptions, we may establish consistency:

Proposition 5 *Suppose that Assumptions 1 and 3 hold and $(\tilde{P}_0,\tilde{\theta}_0)$ is consistent. Suppose we obtain $\tilde{\theta}_k$ by the approximate q-NPL algorithm. Then $\tilde{\theta}_k - \theta^0 = o_p(1)$ for $k = 1,2,\ldots$*

The following proposition establishes that the convergence property of the approximate q-NPL algorithm is the same as that of the original q-NPL algorithm.

Proposition 6 *Suppose Assumptions 1-3 hold and $(\tilde{P}_0,\tilde{\theta}_0)$ is consistent. Suppose we obtain $\{\tilde{P}_j,\tilde{\theta}_j\}_{j=1}^k$ by the approximated q-NPL algorithm. Then, for $j = 1,\ldots,k$,*

$$\begin{aligned} \tilde{\theta}_j - \hat{\theta}_{qNPL} &= O_p(||\tilde{P}_{j-1} - \hat{P}_{qNPL}||), \\ \tilde{P}_j - \hat{P}_{qNPL} &= M_{\Lambda_\theta^q}\Lambda_P^q(\tilde{P}_{j-1} - \hat{P}_{qNPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{qNPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{qNPL}||^2), \end{aligned}$$

where $M_{\Lambda_\theta^q} \equiv I - \Lambda_\theta^q((\Lambda_\theta^q)'\Delta_P\Lambda_\theta^q)^{-1}(\Lambda_\theta^q)'\Delta_P$ with $\Lambda_\theta^q = \nabla_{\theta'}\Lambda^q(P^0,\theta^0)$.

Thus, the approximate q-NPL algorithm achieves the same convergence rate as the original q-NPL algorithm, improving the convergence property of the NPL algorithm if the dominant eigenvalue of Λ_P is less than 1 in absolute value. Upon convergence, this algorithm generates the q-NPL estimator defined by (11), which is more efficient than the NPL estimator $\hat{\theta}_{NPL}$. Thus, even if the NPL algorithm does not encounter any convergence problem, taking additional steps of the approximated q-NPL algorithm starting from $\hat{\theta}_{NPL}$ improves efficiency.

Kasahara and Shimotsu (2006) develop another approximation method based on the Newton-Raphson (NR) algorithm. Given an initial consistent estimator $(\tilde{P}_0,\tilde{\theta}_0)$, this NR-based approximate algorithm iterates

Step 1: Update θ by one NR-step as $\tilde{\theta}_j = \tilde{\theta}_{j-1} - (Q^q_{n,j-1})^{-1}\nabla_{\theta'} L^q_n(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$, where $L^q_n(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ $= n^{-1}\sum_{i=1}^n \ln \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_i|x_i)$ and $Q^q_{n,j-1}$ is a consistent estimator for $\nabla_{\theta\theta'} E[L^q_n(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})]$.

Step 2: Update P using the obtained estimate $\tilde{\theta}_j$ by $\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_j)$

until $j = k$. In Step 1, the parameter θ is updated using one NR step without fully solving the optimization problem. Since taking one NR step brings the estimator sufficiently close to the solution of the original optimization problem, the NR-based approximation algorithm achieves the same rate of convergence as the original q-NPL algorithm. In terms of computation cost, using an outer-product-of-the-gradients estimator for $Q_{n,j-1}$, this algorithm requires the same number of evaluations of $\Lambda^q(P, \theta)$ as the approximate q-NPL algorithm.

The approximate q-NPL algorithm has the following advantages over the NR-based approximate algorithm. First, when the likelihood surface is complex, a simple application of the NR step may not work well in practice. Solving the "linearly approximated likelihood" maximization problem, the approximate q-NPL algorithm essentially applies a version of line-search method without too much of computational cost and is probably more robust than the NR-based approximate algorithm.[8] Second, as we discuss later, we may apply the approximate q-NPL algorithm in the context of the EM algorithm, which is a popular method to estimate models with unobserved heterogeneity (cf., Arcidiacono and Miller (2008)).

4.3 Approximate fixed point algorithm

The approximation method similar to the approximate q-NPL algorithm can be directly applied to the fixed point, $P_\theta = \Psi(P_\theta, \theta)$, resulting in the approximation of the MLE. From Taylor expansion and using $\nabla_{\theta'} P_\theta = (I - \nabla_{P'}\Psi(P_\theta, \theta))^{-1}\nabla_{\theta'}\Psi(P_\theta, \theta)$, we can approximate P_θ as

$$P_\theta = P_{\theta^0} + (I - \nabla_{P'}\Psi(P_{\theta^0}, \theta^0))^{-1}\nabla_{\theta'}\Psi(P_{\theta^0}, \theta^0)(\theta - \theta^0) + O(||\theta - \theta^0||^2), \tag{13}$$

where $\nabla_{\theta'} P_{\theta^0}$ denotes the derivative of P_θ evaluated at $\theta = \theta^0$. Therefore, if we have a consistent estimate of θ^0 and P^0, we may approximate P_θ with the mappings $\nabla_{P'}\Psi(P, \theta)$ and $\nabla_{\theta'}\Psi(P, \theta)$. This approximation method is particularly useful when it is possible to derive an analytical expression for $\nabla_{P'}\Psi(P, \theta)$ and $\nabla_{\theta'}\Psi(P, \theta)$.

Consider the following objective function based on (13):

$$Q_n(\theta, P^*, \theta^*) = n^{-1}\sum_{i=1}^n \ln \Phi(\theta, P^*, \theta^*)(a_i|x_i),$$

where

$$\Phi(\theta, P^*, \theta^*) = P^* + (I - \nabla_{P'}\Psi(P^*, \theta^*))^{-1}\nabla_{\theta'}\Psi(P^*, \theta^*)(\theta - \theta^*). \tag{14}$$

[8] Introducing the line-search method into the NR-based algorithm requires evaluating $\Lambda^q(P, \theta)$ at various step lengths and will substantially increase its computational cost.

We call the estimation algorithm using $Q_n(\theta, P^*, \theta^*)$ the *Approximate Fixed Point Algorithm (AFXP)* because it is based on the approximation of the fixed point P_θ.

Let $\tilde{\theta}_0$ be an initial estimator of θ^0, such as the PML estimator. For $j \geq 1$, consider the following sequential procedure.

Step 1: Given $\tilde{\theta}_{j-1}$, update P by solving the fixed point $P_{\tilde{\theta}_{j-1}} = \Psi(P_{\tilde{\theta}_{j-1}}, \tilde{\theta}_{j-1})$. If there are multiple fixed points, choose the one that maximizes the likelihood:

$$\tilde{P}_j = \underset{P \in \mathcal{M}_{\tilde{\theta}_{j-1}}}{\arg\max} \ln P(a_i|x_i),$$

where $\mathcal{M}_\theta$ is defined in (3).

Step 2: Given $(\tilde{P}_j, \tilde{\theta}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta \in \Theta_j} Q_n(\theta, \tilde{P}_j, \tilde{\theta}_{j-1})$, where

$$\Theta_j = \{\theta \in \Theta : \Phi(\theta, \tilde{\theta}_{j-1}, \tilde{P}_j)(a|x) \in [\varepsilon, 1-\varepsilon] \text{ for all } (a,x) \in A \times X\} \tag{15}$$

for an arbitrary small $\epsilon > 0$.

Iterate Steps 1-2 until $j = k$.

To establish the consistency of sequential estimators generated by the AFXP algorithm, consider the following assumptions. The first set of the assumptions is regularity conditions for the consistency of the MLE. The second set of the assumptions is concerned with the NPL algorithm. Let $\mathcal{N}$ denote a neighborhood of (P^0, θ^0).

Assumption 4 *(a) Θ is compact and, for any $\theta \in \Theta$, $\mathcal{M}_\theta$ is compact. (b) (a_i, x_i) for $i = 1, \ldots, M$, are independently and identically distributed, and* $\Pr(x_i = x) > 0$ *for any $x \in X$. (c) There is a unique $\theta^0 \in$int(Θ) and a unique $P_{\theta^0} \in \mathcal{M}_{\theta^0}$ such that, for any $(a,x) \in A \times X$, $P_{\theta^0}(a|x) = P^0(a|x)$. (d) For any $P_\theta \in \mathcal{M}_\theta$ given any $\theta \neq \theta^0$,* $\Pr_{P^0}(\{(a,x) : P_\theta(a|x) \neq P^0(a|x)\}) > 0$*. (e) $E \sup_{\theta \in \Theta} |P_\theta(a|x)| < \infty$.*

Assumption 5 *(a) $\Psi(P,\theta)(a|x) > 0$ for any $(a,x) \in A \times X$ and any $\{P, \theta\} \in B_P \times \Theta$. (b) $\Psi(P,\theta)$ is continuously differentiable in $(P,\theta) \in \mathcal{N}$, and* $\sup_{(P,\theta)\in\mathcal{N}} ||\nabla_{P'}\Psi(P,\theta)|| < \infty$ *and* $\sup_{(P,\theta)\in\mathcal{N}} ||\nabla_{\theta'}\Psi(P,\theta)|| < \infty$*.*

The consistency of AFXP estimator requires the following additional assumptions:

Assumption 6 *(a) For any $\eta \in \mathbb{R}^K$ such that $\eta \neq 0$, $\nabla_{\theta'} P_{\theta^0}(a|x)\eta \neq 0$ with positive probability. (b) $E \sup_{\theta\in\Theta,(P^*,\theta^*)\in\mathcal{N}} ||\nabla_{\theta^{*\prime}}\Phi(\theta, P^*, \theta^*)(a|x)|| < \infty$, and $E \sup_{\theta\in\Theta,(P^*,\theta^*)\in\mathcal{N}} ||\nabla_{P^{*\prime}}\Phi(\theta, P^*, \theta^*)(a|x)|| < \infty$. (c) $E||\nabla_{\theta'} P_{\theta^0}(a|x)|| < \infty$.*

Assumption 6(a) is similar to Assumption 3 and is an identification condition for the probability limit of our objective function. Assumption 6(b)-(c) are regularity conditions required for the uniform convergence of the objective function. Assumption 6(b) is stated in terms of the conditions on the derivatives of Φ to simplify the presentation, but it is possible to state it in terms of the conditions on the derivatives of $\Psi(P,\theta)$.

Under these assumptions, the sequential estimators generated by the AFXP algorithm is consistent:

Proposition 7 *Suppose that Assumptions 4-6 hold and $\tilde{\theta}_0$ is consistent. Suppose we obtain $\tilde{\theta}_k$ by the AFXP algorithm. Then $\tilde{\theta}_k - \theta^0 = o_p(1)$ for $k = 1, 2, \ldots$*

If a sequence of estimators generated by the AFXP algorithm converges, it converges to the MLE. We now analyze the convergence property of the AFXP algorithm. We introduce the following additional regularity conditions. Let $\mathcal{N}_1$ denote a neighborhood of (P^0, θ^0), and let $\mathcal{N}_2$ denote a neighborhood of (θ^0). Let $\nabla^{(3)}\Phi(\theta, P^*, \theta^*)$ denote the third derivatives of $\Phi(\theta, P^*, \theta^*)$ with respect to (θ, P^*, θ^*). Assumption 7(a) is required for the asymptotic normality of the NFXP estimator.

Assumption 7 *(a) $E\sup_{\theta\in\mathcal{N}_2}||\nabla_{\theta'}P_\theta(a|x)||^2 < \infty$, and $E\sup_{\theta\in\mathcal{N}_2}||\nabla_{\theta\theta'}P_\theta(a|x)|| < \infty$. (b) $\Psi(P,\theta)$ is twice continuously differentiable in $(P,\theta)\in\mathcal{N}_1$ with a bounded second derivative. (c) $E\sup_{\theta\in\mathcal{N}_2,(P^*,\theta^*)\in\mathcal{N}_1}||\nabla^{(3)}\Phi(\theta,P^*,\theta^*)(a|x)|| < \infty$.*

The following proposition establishes the convergence rate of the AFXP algorithm. Let θ_{MLE} be the MLE of θ and define $\hat{P}_{MLE} = P_{\hat{\theta}_{MLE}}$, the MLE of P.

Proposition 8 *Suppose that Assumptions 4-7 hold and $\tilde{\theta}_0$ is consistent. Suppose we obtain $\{\tilde{P}_j, \tilde{\theta}_j\}_{j=1}^k$ by the AFXP algorithm. Then, for $j = 1, 2, \ldots, k$,*

$$\begin{aligned}\tilde{\theta}_j - \hat{\theta}_{MLE} &= O_p(||\tilde{P}_j - \hat{P}_{MLE}||),\\ \tilde{P}_j - \hat{P}_{MLE} &= O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{MLE}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{MLE}||^2).\end{aligned}$$

Thus, the estimator generated by the AFXP algorithm is first-order equivalent to the MLE for all $k \geq 1$. This convergence rate is also the same as that of the NPL algorithm for a single-agent model with $\nabla_P\Psi(P^0,\theta^0) = 0$ (Kasahara and Shimotsu, 2006). This algorithm can be used to obtain the MLE because, upon convergence, its limit is identical to the MLE.

Implementing Step 1 of the AFXP algorithm may be impractical when finding all the fixed points is computationally infeasible. In such cases, we may replace the solution to the fixed point in Step 1 with its consistent estimator as follows. Let $(\tilde{P}_0, \tilde{\theta}_0)$ be an initial estimator of (P^0, θ^0). For $j \geq 1$, consider the *q-AFXP* algorithm which iterates

Step 1: Given $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$, update P by $\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$, and

Step 2: Given $(\tilde{P}_j, \tilde{\theta}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta \in \Theta_j} Q_n(\theta, \tilde{P}_j, \tilde{\theta}_{j-1})$, where Θ_j is given by (15),

until $j = k$. The sequential estimators generated by the q-AFXP algorithm is consistent.

Proposition 9 *Suppose that Assumptions 4-6 hold and $(\tilde{P}_0, \tilde{\theta}_0)$ is consistent. Suppose we obtain $\tilde{\theta}_k$ by the q-AFXP algorithm. Then $\tilde{\theta}_k - \theta^0 = o_p(1)$ for $k = 1, 2, \ldots$*

We now derive the convergence property of the q-AFXP algorithm. First, we introduce some notations. Define the information matrix for the MLE as $\mathcal{I}^0 = E[\nabla_\theta \ln P_{\theta^0}(a_i|x_i) \nabla_{\theta'} \ln P_{\theta^0}(a_i|x_i)]$. Under the standard regularity conditions, the MLE satisfies $\sqrt{n}(\hat{\theta}_{MLE} - \theta^0) \to_d N(0, (\mathcal{I}^0)^{-1})$. Define a $K \times L$ matrix J as (we state it in terms of J' for notational convenience)

$$J' = E\left[\nabla_P \left\{ \frac{[(I - \nabla_{P'}\Psi(P^0, \theta^0))^{-1} \nabla_{\theta'} \Psi(P^0, \theta^0)](a|x)}{P^0(a|x)} \right\}\right].$$

The following proposition establishes the convergence rate of the q-AFXP algorithm.

Proposition 10 *Suppose that Assumptions 4-7 hold and $(\tilde{P}_0, \tilde{\theta}_0)$ is consistent. Suppose we obtain $\tilde{\theta}_k$ by the q-AFXP algorithm. Then, for $k = 1, 2, \ldots$,*

$$\begin{aligned} \tilde{P}_j - \hat{P}_{MLE} &= \Lambda_P^q (\tilde{P}_{j-1} - \hat{P}_{MLE}) + \Lambda_\theta^q (\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) + R_{n,j}, \\ (\mathcal{I}^0 + o_p(1))(\tilde{\theta}_j - \hat{\theta}_{MLE}) &= -J \nabla_{\theta'} P_{\theta^0} (\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) + J(\tilde{P}_j - \hat{P}_{MLE}) + R_{n,j}, \end{aligned}$$

where $R_{n,j}$ denotes a generic reminder term satisfying

$$\begin{aligned} R_{n,j} &= O_p(||\tilde{P}_{j-1} - \hat{P}_{MLE}||^2) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) \\ &\quad + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{MLE}||) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||). \end{aligned}$$

Ignoring $R_{n,j}$ and $o_p(1)$ term and arranging the two updating relations into a system of equations, we obtain

$$\begin{pmatrix} I_L & 0 \\ -J & \mathcal{I}^0 \end{pmatrix} \begin{pmatrix} \tilde{P}_j - \hat{P}_{MLE} \\ \tilde{\theta}_j - \hat{\theta}_{MLE} \end{pmatrix} = \begin{pmatrix} \Lambda_P^q & \Lambda_\theta^q \\ 0 & -J\nabla_{\theta'} P_{\theta^0} \end{pmatrix} \begin{pmatrix} \tilde{P}_{j-1} - \hat{P}_{MLE} \\ \tilde{\theta}_{j-1} - \hat{\theta}_{MLE} \end{pmatrix}.$$

It follows that

$$\begin{pmatrix} \tilde{P}_j - \hat{P}_{MLE} \\ \tilde{\theta}_j - \hat{\theta}_{MLE} \end{pmatrix} = Q \begin{pmatrix} \tilde{P}_{j-1} - \hat{P}_{MLE} \\ \tilde{\theta}_{j-1} - \hat{\theta}_{MLE} \end{pmatrix}, \text{ where } Q = \begin{pmatrix} \Lambda_P^q & \Lambda_\theta^q \\ (\mathcal{I}^0)^{-1} J \Lambda_P^q & (\mathcal{I}^0)^{-1} J (\Lambda_\theta^q - \nabla_{\theta'} P_{\theta^0}) \end{pmatrix}.$$

Therefore, the convergence property of the q-AFXP algorithm, or the eigenvalues of Q, depends on three factors: (i) the magnitude of Λ_P^q and Λ_θ^q, (ii) the magnitude of $\mathcal{I}^0$ and J, and

(iii) difference between Λ_θ^q and $\nabla_{\theta'} P_{\theta^0}$. From the properties of Λ, we obtain $\Lambda_P^q = (\Lambda_P)^q$ and $\Lambda_\theta^q - \nabla_{\theta'} P_{\theta^0} = -\nabla_{\theta'} P_{\theta^0} (\Lambda_P)^{q-1}$. Since the trace of a matrix is the sum of its eigenvalues, the sum of the eigenvalues of Q is the sum of the trace of $(\Lambda_P)^q$ and the trace of $-(\mathcal{I}^0)^{-1} J \nabla_{\theta'} P_{\theta^0} (\Lambda_P)^{q-1}$. Therefore, when all the eigenvalues of Λ_P are smaller than 1 in absolute value, then, for sufficiently large q, all the eigenvalues of Q are smaller than 1, and iterating the q-AFXP algorithm converges to the MLE.

5 Sequential GMM estimators

Recently, many researchers extend the Hotz-Miller CCP estimator and develop various two-step moment estimators for dynamic games (see Bajari, Benkard and Levin (2007), Pakes, Ostrovsky and Berry (2007), Pesendorfer and Schmidt-Dengler (2007)). The main advantages of these two step moment estimators are, first, their computational simplicity and, second, their consistency property given the consistent estimator for P^0 even when the mapping $\Psi(\cdot, \theta^0)$ has multiple fixed points (i.e., multiple equilibria). These estimators often suffer from the finite sample bias, however, especially when the initial estimator for P^0 is imprecise.

This section develops a recursive extension of two-step moment estimators called the *sequential GMM estimator* using the similar idea to the NPL algorithm. The sequential GMM estimator is asymptotically more efficient and may have a smaller finite sample bias than two-step moment estimators.

5.1 GMM estimator

Given the conditional probabilities P^0 in population, for any function $h : A \to \mathbb{R}$, the following conditional moment condition always holds: $E\left[h(a) - \sum_{a' \in A} h(a') P^0(a'|x) \mid x\right] = 0$. For example, we may choose $h(a) = a$ or $h(a) = a^2$. The conditional moment condition imply unconditional moment conditions of the form

$$E\left[g_l(x, a; P^0)\right] = 0,$$

where

$$g_l(x, a; P^0) = \rho_l(x) \left(h_l(a) - \sum_{a' \in A} h_l(a') P^0(a'|x) \right) \tag{16}$$

for any function $\rho_l : X \to \mathbb{R}$ and $h_l : A \to \mathbb{R}$.

We consider the generalized method of moments estimator based on these moment conditions when the population conditional probabilities belong to a parametric class of conditional probabilities with fixed point constraint: $\mathcal{M} = \cup_{\theta \in \Theta} \mathcal{M}_\theta = \{P \in B_P : \ P = \Psi(P, \theta), \ \theta \in \Theta\}$.

The generalized method of moments (GMM) estimator with fixed point constraint is defined as:

$$\hat{\theta}_{GMM} = \arg\min_{\theta \in \Theta} \left\{ \min_{P \in \mathcal{M}_\theta} \ \bar{g}(P)'\hat{W}\bar{g}(P) \right\}, \tag{17}$$

where $\mathcal{M}_\theta$ is given in (3), $\hat{W} \to_p W$ positive semi-definite, and

$$\bar{g}(P) = n^{-1}\sum_{i=1}^{n} g(a_i, x_i; P),$$

where $g(\cdot; P) = (g_1(\cdot; P), g_2(\cdot; P), ..., g_L(\cdot; P))'$ is a moment vector function representing L moment conditions with $g_l(\cdot)$ function given by (16) for $l = 1, .., L$.

To compute the GMM estimator, we need to repeatedly solve the fixed point of $P = \Psi(P, \theta)$ for each candidate parameter value θ until one finds the parameter that minimizes the GMM objective function. When solving the fixed point is costly as in models of dynamic game, this estimator is impractical.

5.2 Two-step GMM estimator

The two-step GMM estimator is defined as

$$\hat{\theta}_{2GMM} \quad = \quad \arg\min_{\theta \in \Theta} \ \bar{g}(\Psi(\hat{P}_0, \theta))'\hat{W}\bar{g}(\Psi(\hat{P}_0, \theta)),$$

where $\hat{P}_0$ is an initial consistent estimator for P^0.

In the following, we use the following notations.

$$\begin{aligned} \bar{G}_\theta(\Psi(P, \theta)) = (\partial/\partial\theta')\bar{g}(\Psi(P, \theta)), \qquad & \bar{G}_P(\Psi(P, \theta)) = (\partial/\partial P')\bar{g}(\Psi(P, \theta)), \\ G_\theta = E[(\partial/\partial\theta')g(a_i, x_i; \Psi(P^0, \theta^0))], \qquad & G_P = E[(\partial/\partial P')g(a_i, x_i; \Psi(P^0, \theta^0))]. \end{aligned}$$

Define f_x as before so that its elements are arranged conformably with $P^0(j|x^l)$ while let $\hat{f}_x$ be a frequency estimator of f_x. Denote $\Delta_x = diag(f_x)$ and $\hat{\Delta}_x = diag(\hat{f}_x)$. Let $\gamma_l(a, x) = \rho_l(x)h_l(a)$ and γ_l represent a vector of $|A||X|$ length. Finally, let $\Gamma = (\gamma_1', \gamma_2', ..., \gamma_L')'$ be a L by $|A||X|$ matrix. With those notations, we may write $\bar{G}_\theta(\Psi(P, \theta)) = -\Gamma\hat{\Delta}_x(\partial/\partial\theta')\Psi(P, \theta)$, $\bar{G}_P(\Psi(P, \theta)) = -\Gamma\hat{\Delta}_x(\partial/\partial P')\Psi(P, \theta)$, $G_\theta = -\Gamma\Delta_x\Psi_\theta$ and $G_P = -\Gamma\Delta_x\Psi_P$. Let $r(a_i, x_i)$ be a vector of length $|A||X|$ whose elements are arranged conformably with $P^0(a|x)$ and be equal to zero except for the element of $(a, x) = (a_i, x_i)$ which takes a value of one. With this notation, we can write $\hat{P}_0 = n^{-1}\sum_{i=1}^{n} r(a_i, x_i)$.

Assumption 8 *(a) For any $\theta \neq \theta^0$, $WE[g(a, x; \Psi(P^0, \theta))] \neq 0$; (b) $G_\theta' W G_\theta$ is nonsingluar; (c) $E[||g(a, x; P^0)||^2] < \infty$; (d) $E[\sup_{\theta \in \Theta} ||g(a, x; \Psi(P^0, \theta))||] < \infty$; (e) $E[\sup_{\theta \in \Theta} ||\nabla_{\theta'} g(a, x; \Psi(P^0, \theta))||] <$*

∞.

Under Assumptions 1 and 8, $\hat{\theta}_{2GMM}$ is consistent and asymptotic normal. The asymptotic distribution of $\hat{\theta}_{2GMM}$ is given by

$$\sqrt{n}(\hat{\theta}_{2GMM} - \theta^0) \to_d N(0, V_{2GMM}),$$

where $V_{2GMM} = (G_\theta' W G_\theta)^{-1} G_\theta' W S W G_\theta (G_\theta' W G_\theta)^{-1}$ with $S = E[(g(a_i, x_i; P^0) + G_P(r(a_i, x_i) - P^0))(g(a_i, x_i; P^0) + G_P(r(a_i, x_i) - P^0))']$. Using an optimal weighting matrix $W = S^{-1}$, the limiting variance is given by $V_{2GMM} = (G_\theta' S^{-1} G_\theta)^{-1}$.

5.3 The nested GMM estimator

We consider a recursive extension of the two-step GMM estimator called the *nested GMM algorithm* as follows. Given an initial estimator $\tilde{P}_0$,

Step 1: Given $\tilde{P}_{j-1}$, update θ by $\tilde{\theta}_j = \arg\min_\theta \bar{g}(\Psi(\tilde{P}_{j-1}, \theta))' \hat{W} \bar{g}(\Psi(\tilde{P}_{j-1}, \theta))$.

Step 2: Update P using the obtained estimate $\tilde{\theta}_j$: $\tilde{P}_j = \Psi(\tilde{P}_{j-1}, \tilde{\theta}_j)$.

Iterate Steps 1-2 until $j = k$.

Under regularity conditions similar to the ones in Assumption 1, the sequence of estimators generated by this algorithm are consistent. If the sequence converges, the limit is called the nested GMM (NGMM) estimator. The NGMM estimator $(\hat{P}_{NGMM}, \hat{\theta}_{NGMM})$ satisfies

$$\begin{aligned}
\hat{\theta}_{NGMM} &= \arg\min_{\theta\in\Theta} \quad \bar{g}(\Psi(\hat{P}_{NGMM}, \theta))' \hat{W} \bar{g}(\Psi(\hat{P}_{NGMM}, \theta)), \\
\hat{P}_{NGMM} &= \Psi(\hat{P}_{NGMM}, \hat{\theta}_{NGMM}).
\end{aligned}$$

Under the following additional assumptions, we derive the limiting distribution of the NGMM estimator.

Assumption 9

$$\begin{aligned}
&\bar{g}(P^0) = O_p(n^{-1/2}), \quad \sup_{\theta,P} ||D^2\Psi(P,\theta)|| < \infty, \quad ||\Gamma|| < \infty, \\
&\mathit{rank}((\partial/\partial\theta')\Psi(P,\theta)) = k, \ \mathit{for\ all}\ P
\end{aligned}$$

Note that $\sup_{\theta,P} ||D^2\Psi(P,\theta)|| < \infty$ and $||\Gamma|| < \infty$ imply that $\sup_{\theta,P} ||D\bar{G}_\theta(\Psi(P,\theta))|| < \infty$. The rank condition on $(\partial/\partial\theta')\Psi(P,\theta)$ guarantees that $(\bar{G}_\theta(P))' \hat{W} \bar{G}_\theta(P)$ is invertible.

Proposition 11 *Suppose Assumptions 1, 8 and 9 hold. Then*

$$\sqrt{n}(\hat{\theta}_{NGMM} - \theta^0) \to_d N(0, (G_\theta' W G_\theta^\infty)^{-1} G_\theta' W \Omega W' G_\theta ((G_\theta^\infty)' W' G_\theta)^{-1}),$$

where $\Omega = E[g(a_i, x_i; P^0)g(a_i, x_i; P^0)']$ and $G_\theta^\infty = -\Gamma\Delta_x(I - \Psi_P)^{-1}\Psi_\theta$. If we choose $W = \Omega^{-1}$, the asymptotic variance is given by $(G_\theta'\Omega^{-1}G_\theta^\infty)^{-1}G_\theta'\Omega^{-1}G_\theta((G_\theta^\infty)'\Omega^{-1}G_\theta)^{-1}$.

Remark 6 *When $\Psi_P = 0$, the two-step GMM estimator with an optimal weighting matrix is asymptotically equivalent to the NGMM estimator with $W = \Omega^{-1}$.*

The NGMM estimator can be obtained as the limit of the sequence of estimators generated by the NGMM algorithm upon convergence. The convergence property of the NGMM estimator is given by the following lemma.

Proposition 12 *Suppose Assumptions 1 and 9 hold. Then, for $j = 1, \dots, k$,*

$$\begin{aligned}
\tilde{\theta}_j - \tilde{\theta} &= O_p(||\tilde{P}_{j-1} - \tilde{P}||),\\
\tilde{P}_j - \tilde{P} &= [I + \Psi_\theta(G_\theta'\hat{W}G_\theta)^{-1}G_\theta'\hat{W}\Gamma\Delta_x]\Psi_P(\tilde{P}_{j-1} - \tilde{P}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2).
\end{aligned}$$

Remark 7 *Observe that $-\Psi_\theta(G_\theta'\hat{W}G_\theta)^{-1}G_\theta'\hat{W}\Gamma\Delta_x = \Psi_\theta(\Psi_\theta'\Delta_x'\Gamma'\hat{W}\Gamma\Delta_x\Psi_\theta)^{-1}\Psi_\theta'\Delta_x'\Gamma'\hat{W}\Gamma\Delta_x$ is a projection matrix, and the sequence of estimators here also has the convergence property similar to the estimators generated by the NPL algorithm. Again, the convergence rate is primarily determined by the eigenvalues of Ψ_P.*

Remark 8 *Analogous remarks to Remarks 1-5 apply here.*

5.4 Alternative sequential GMM estimators

When the NGMM algorithm encounters a convergence problem or when researchers are interested in obtaining more efficient GMM estimator, we may use alternative sequential GMM algorithms that can be developed in the same spirit as those of Section 4.

First, as in the case of the NPL algorithm, using Λ in place of Ψ in the NGMM algorithm improves the convergence property while both share the same limit.

Second, replacing Ψ with Λ^q in the NGMM algorithm improves not only convergence property but also may increase the asymptotic efficiency if the dominant eigenvalue of Λ is less than 1. The NGMM estimator with Λ^q converges to the GMM estimator (17) as $q \to \infty$. To reduce computational burden, given an initial consistent estimator $(\tilde{P}_0, \tilde{\theta}_0)$, we may approximate the mapping $\Lambda^q(P, \theta) \approx \Lambda^q(P_{j-1}, \theta_{j-1}) + \nabla_{\theta'}\Lambda^q(P_{j-1}, \theta_{j-1})(\theta - \tilde{\theta}_{j-1})$ in Step 1 of the NGMM algorithm to obtain the approximate q-NGMM algorithm with Λ^q which iterates

Step 1: Given $\tilde{P}_{j-1}$, update θ by $\tilde{\theta}_j = \arg\min_\theta \bar{g}(\tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}))'\hat{W}\bar{g}(\tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}))$, where $\tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \equiv \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta'}\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\theta - \tilde{\theta}_{j-1})$, and

Step 2: Update P using the obtained estimate $\tilde{\theta}_j$: $\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_j)$.

If the iterations converge, the limit of the sequence generated by the approximate q-NGMM algorithm is the same as the NGMM estimator with Λ^q.

Finally, the approximate GMM (AGMM) algorithm can be defined analogously to the AFXP algorithm as follows. Given an initial consistent estimator $(\tilde{P}_0, \tilde{\theta}_0)$, iterate

Step 1: Given $\tilde{\theta}_{j-1}$, update P by solving the fixed point $P_{\tilde{\theta}_{j-1}} = \Psi(P_{\tilde{\theta}_{j-1}}, \tilde{\theta}_{j-1})$. If there are multiple fixed points, update P by $\tilde{P}_j = \arg\max_{P \in \mathcal{M}_{\tilde{\theta}_{j-1}}} \bar{g}(P)'\hat{W}\bar{g}(P)$.

Step 2: Given $(\tilde{P}_j, \tilde{\theta}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\min_{\theta\in\Theta} \left\{ \min_{P \in \mathcal{M}_\theta} \ \bar{g}(\Phi(\theta, \tilde{\theta}_{j-1}, \tilde{P}_j))'\hat{W}\bar{g}(\Phi(\theta, \tilde{\theta}_{j-1}, \tilde{P}_j)) \right\}$, where $\Phi(\theta, P^*, \theta^*)$ and Θ_j are defined by (14) and (15), respectively.

When solving the fixed point is computationally difficult in Step 1, one may update P by $\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ in Step 1 to define the q-AGMM algorithm in the same spirit as the q-AFXP algorithm and, upon convergence, the GMM estimator (17) can be obtained.

6 Unobserved Heterogeneity

This section extends our analysis to models with unobserved heterogeneity. The NPL algorithm has important advantage over two step methods in estimating models with unobserved heterogeneity because obtaining a reliable initial estimate of P is difficult in this context.

6.1 Permanent unobserved heterogeneity

Suppose that there are M types of agents, where type m is characterized by a type-specific parameter θ^m and the population probability of being type m is π^m with $\sum_{m=1}^M \pi^m = 1$. These types capture time-invariant state variables that are unobserved by researchers. With a slight abuse of notation, denote $\theta = (\theta^1, ..., \theta^M)' \in \Theta^M$ and $\pi = (\pi^1, ..., \pi^M)' \in \Theta_\pi$. Then, $\zeta = (\theta', \pi')'$ is the parameter to be estimated, and let $\Theta_\zeta = \Theta^M \times \Theta_\pi$ denote the set of possible values of ζ. The true parameter is denoted by ζ^0.

Consider a panel data set $\{\{a_{it}, x_{it}, x_{i,t+1}\}_{t=1}^T\}_{i=1}^n$ such that $w_i = \{a_{it}, x_{it}, x_{i,t+1}\}_{t=1}^T$ is randomly drawn across i's from the population. The conditional probability distribution of a_{it} given x_{it} for type m agent is given by a fixed point of $P_{\theta^m} = \Psi(P_{\theta^m}, \theta^m)$. On the other hand, to simplify our analysis, we assume that the transition probability function of x_{it} is independent of types and given by $f_x(x_{i,t+1}|a_{it}, x_{it})$ and is known to researchers.[9]

In this framework, the initial state x_{i1} is correlated with unobserved type (i.e., the initial conditions problem of Heckman (1981)). We assume that x_{i1} for type m is randomly drawn

[9] When the transition probability function is independent of types, it can be directly estimated from transition data without solving the fixed point problem. Kasahara and Shimotsu (2006) analyze the case in which the transition probability function is also type-dependent in the context of a single agent dynamic programming model with unobserved heterogeneity.

from the type m stationary distribution characterized by a fixed point of the following equation: $p^*(x) = \sum_{x' \in X} p^*(x') \left(\sum_{a' \in A} P_{\theta^m}(a'|x') f_x(x|a', x') \right) \equiv [T(p^*, P_{\theta^m})](x)$. Since solving the fixed point of $T(\cdot, P)$ for given P is often less computationally intensive than computing the fixed point of $\Psi(\cdot, \theta)$, we assume the full solution of the fixed point of $T(\cdot, P)$ is available given P.

Stack P^m's as $\mathbf{P} = (P^{1'}, \ldots, P^{M'})'$, and let $\mathbf{P}^0$ denote its true value. Define $\mathbf{\Psi}(\mathbf{P}, \theta) = (\Psi(P^1, \theta^1)', \ldots, \Psi(P^M, \theta^M)')'$. Then, the set of possible probabilities consistent with the fixed point constraints given the value of θ is given by $\mathcal{M}_\theta^* = \{\mathbf{P} \in B_P^M : \ \mathbf{P} = \mathbf{\Psi}(\mathbf{P}, \theta)\}$.

The maximum likelihood estimator for a model with unobserved heterogeneity is:

$$\hat{\zeta}_{MLE} = \arg\max_{\zeta \in \Theta_\zeta} \left\{ \max_{\mathbf{P} \in \mathcal{M}_\theta^*} \ln \left([L(\mathbf{P}, \pi)](w_i) \right) \right\}, \tag{18}$$

where

$$[L(\mathbf{P}, \pi)](w_i) = \sum_{m=1}^{M} \pi^m p_{P^m}^*(x_{i1}) \prod_{t=1}^{T} P^m(a_{it}|x_{it}) f_x(x_{i,t+1}|a_{it}, x_{it}),$$

and $p_{P^m}^* = T(p_{P^m}^*, P^m)$ is the type m stationary distribution of x when the conditional probability is P^m. If $\mathbf{P}^0$ is the true conditional probability distribution and π^0 is the true mixing distribution, then $L^0 = L(\mathbf{P}^0, \pi^0)$ represents the true probability distribution of w.

Assumption 10 *(a) Θ_ζ is compact and, for any $\theta \in \Theta^M$, $\mathcal{M}_\theta^*$ is compact. (b) $[L(\mathbf{P}, \pi)](w) > 0$ for any w and for any $(\mathbf{P}, \pi) \in \cup_{\theta \in \Theta^M} \mathcal{M}_\theta^* \times \Theta_\pi$. (c) $w_i = \{(a_{it}, x_{it}, x_{i,t+1}) : t = 1, \ldots, T\}$ for $i = 1, \ldots, n$, are independently and identically distributed. (d) For any $P \in B_P$, there exists a unique fixed point for $T(\cdot, P)$. (e) There is a unique $\zeta^0 \in int(\Theta_\zeta)$ and a unique $\mathbf{P}^0 \in \mathcal{M}_{\theta^0}^*$ such that, for any $w = \{(a_t, x_t, x_{t+1}) : t = 1, \ldots, T\}$, $[L(\mathbf{P}^0, \pi^0)](w) = L^0(w)$, where L^0 is the true probability for w. Given any $(\theta, \pi) \neq (\theta^0, \pi^0)$, for any $\mathbf{P} \in \mathcal{M}_\theta^*$, $\Pr_{L^0}(\{w : [L(P, \pi)](w) \neq L^0(w)\}) > 0$. (f) $\tilde{\mathbf{P}}_0 - \mathbf{P}^0 = o_p(1)$, and the MLE denoted by $\hat{\zeta}_{MLE}$ satisfies $\sqrt{n}(\hat{\zeta}_{MLE} - \zeta^0) \rightarrow_d N(0, \Omega_\zeta)$.*

Assumption 10(f) requires an initial consistent estimators for the type-specific conditional probabilities. Kasahara and Shimotsu (2006) derive sufficient conditions for nonparametric identification of a finite mixture model and suggest a sieve estimator which can be used to obtain an initial consistent estimate for $\mathbf{P}$. On the other hand, as AM07 argue, if the NPL algorithm converges, then the limit may provide a consistent estimate for the parameter ζ even when $\tilde{\mathbf{P}}_0$ is not consistent.

We consider a version of the NPL algorithm for a model with unobserved heterogeneity originally developed by AM07 as follows. Assume that an initial consistent estimator $\tilde{\mathbf{P}}_0 = (\tilde{P}_0^1, \ldots, \tilde{P}_0^M)$ is available. For $j = 1, 2, \ldots$, iterate

Step 1: Given $\tilde{\mathbf{P}}_{j-1}$, update $\zeta = (\theta', \pi')'$ by

$$\tilde{\zeta}_j = \arg\max_{\zeta \in \Theta_\zeta} \ n^{-1} \sum_{i=1}^{n} \ln \left([L(\mathbf{\Psi}(\tilde{\mathbf{P}}_{j-1}, \theta), \pi)](w_i) \right),$$

Step 2: Update $\mathbf{P}$ using the obtained estimate $\tilde{\theta}_j$ by $\tilde{\mathbf{P}}_j = \boldsymbol{\Psi}(\tilde{\mathbf{P}}_{j-1}, \tilde{\theta}_j)$,

until $j = k$. If the iterations converge, its limit $(\hat{\mathbf{P}}_{NPL}, \hat{\zeta}_{NPL})$ is the NPL estimator for models with unobserved heterogeneity and satisfies the conditions analogous to (9).

We now establish the convergence property of the NPL algorithm for models with unobserved heterogeneity.

Let $[l(\mathbf{P},\zeta)](w) \equiv \ln\Big([L(\boldsymbol{\Psi}(\tilde{\mathbf{P}}_{j-1},\theta),\pi)](w)\Big)$. Then, $\Omega_{\zeta\zeta} = E\left[(\partial/\partial\zeta)[l(\mathbf{P}^0,\zeta^0)](w)(\partial/\partial\zeta')[l(\mathbf{P}^0,\zeta^0)](w)\right]$ and $\Omega_{\zeta P} = E\left[(\partial/\partial\zeta)[l(\mathbf{P}^0,\zeta^0)](w)(\partial/\partial\mathbf{P}')[l(\mathbf{P}^0,\zeta^0)](w)\right]$ can be written as

$$\Omega_{\zeta\zeta} = \begin{bmatrix} \Omega_{\theta\theta} & \Omega_{\theta\pi} \\ \Omega_{\pi\theta} & \Omega_{\pi\pi} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Psi}_\theta' L_P' \Delta_L L_P \boldsymbol{\Psi}_\theta & \boldsymbol{\Psi}_\theta' L_P' \Delta_L L_\pi \\ L_\pi' \Delta_L L_P \boldsymbol{\Psi}_\theta & L_\pi' \Delta_L L_\pi \end{bmatrix}, \quad \Omega_{\zeta P} = \begin{bmatrix} \Omega_{\theta P} \\ \Omega_{\pi P} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Psi}_\theta' L_P' \Delta_L L_P \boldsymbol{\Psi}_P \\ L_\pi' \Delta_L L_P \boldsymbol{\Psi}_P \end{bmatrix},$$

where $\boldsymbol{\Psi}_P \equiv (\partial/\partial\mathbf{P}')\boldsymbol{\Psi}(\mathbf{P}^0,\theta^0)$, $\boldsymbol{\Psi}_\theta \equiv (\partial/\partial\theta')\boldsymbol{\Psi}(\mathbf{P}^0,\theta^0)$, $\Delta_L = diag((L^0)^{-1}) = diag((L(\mathbf{P}^0,\pi^0))^{-1})$, $L_P = \nabla_{P'} L(\mathbf{P}^0,\pi^0)$, and $L_\pi = \nabla_{\pi'} L(\mathbf{P}^0,\pi^0)$.

Assumption 11

$$\begin{aligned} \bar{l}_\zeta(\mathbf{P}^0,\zeta^0) &= O_p(n^{-1/2}), & \bar{l}_{\zeta\zeta}(\mathbf{P}^0,\zeta^0) &= -\Omega_{\zeta\zeta} + O_p(n^{-1/2}) \\ \bar{l}_{\zeta P}(\mathbf{P}^0,\zeta^0) &= -\Omega_{\zeta P} + O_p(n^{-1/2}), & \bar{l}_{\zeta\zeta}(P,\theta) & \textit{ is invertible for all } (P,\theta). \\ E\sup_{\zeta,\mathbf{P}} ||D_{\zeta\mathbf{P}}[l(\mathbf{P},\zeta)](w)|| &< \infty, & E\sup_{\zeta,\mathbf{P}} ||D^3[l(\mathbf{P},\zeta)](w)|| &< \infty, \\ \sup_{\theta,P} ||D^2\Psi(P,\theta)|| &= O(1), \end{aligned}$$

where $\bar{l}_\zeta(\mathbf{P},\zeta) = n^{-1}\sum_{i=1}^n (\partial/\partial\zeta)[l(\mathbf{P},\zeta)](w_i)$, $\bar{l}_{\zeta\zeta}(\mathbf{P},\zeta) = n^{-1}\sum_{i=1}^n (\partial^2/\partial\zeta\partial\zeta')[l(\mathbf{P},\zeta)](w_i)$, *and* $\bar{l}_{\zeta\mathbf{P}}(\mathbf{P},\zeta) = n^{-1}\sum_{i=1}^n (\partial^2/\partial\zeta\partial\mathbf{P}')[l(\mathbf{P},\zeta)](w_i)$.

The following result states the convergence properties of the NPL algorithm for models with unobserved heterogeneity.

Lemma 2 *Suppose Assumptions 10-11 hold. Then, for* $j = 1, \ldots, k$,

$$\begin{aligned} \tilde{\zeta}_j - \hat{\zeta}_{NPL} &= O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||), \\ \tilde{\mathbf{P}}_j - \hat{\mathbf{P}}_{NPL} &= [I - \boldsymbol{\Psi}_\theta D \boldsymbol{\Psi}_\theta' L_P' \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P]\boldsymbol{\Psi}_P(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) \\ &\quad + O_p(n^{-1/2}||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||) + O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||^2). \end{aligned}$$

where $D = (\boldsymbol{\Psi}_\theta' L_P' \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P \boldsymbol{\Psi}_\theta)^{-1}$ *and* $M_{L_\pi} = I - \Delta_L^{1/2} L_\pi (L_\pi' \Delta_L L_\pi)^{-1} L_\pi \Delta_L^{1/2}$.

Since $I - \boldsymbol{\Psi}_\theta D \boldsymbol{\Psi}_\theta' L_P' \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P$ is an idempotent matrix, its eigenvalues are either zero or one. Consequently, the dominant eigenvalue of $\boldsymbol{\Psi}_P$ determines the convergence rate of the NPL algorithm for models with unobserved heterogeneity. When the NPL algorithm encounters a convergence problem, replacing $\Psi(P,\theta)$ with $\Lambda(P,\theta)$ improves convergence property.

Remark 9 *It is possible to relax the stationarity assumption on the initial states by estimating the type-specific initial distributions of x, denoted by $\{p^{*m}\}_{m=1}^{M}$, without imposing stationarity restriction in Step 1 of the NPL algorithm. In this case, the NPL algorithm has the convergence rates similar to those of Lemma 2.*

Define $\zeta^* = (\theta', \pi', p^{*'})$ where $p^* = (p^{*1'}, ..., p^{*M'})'$. Define $[L(\mathbf{P}, \pi, p^*)](w_i) = \sum_{m=1}^{M} \pi^m p^{*m}(x_{i1}) \prod_{t=1}^{T} P^m(a_{it}|x_{it}) f_x(x_{i,t+1}|a_{it}, x_{it})$. In this case, the NPL algorithm updates $\tilde{\zeta}^*$ in Step 1 as $\tilde{\zeta}_j^* = \arg\max_{\zeta^* \in \Theta_{\zeta^*}} \ n^{-1} \sum_{i=1}^{n} \ln \left([L(\mathbf{\Psi}(\tilde{\mathbf{P}}_{j-1}, \theta), \pi, p^*)](w_i) \right)$, where $\Theta_{\zeta^*} = \Theta_\zeta \times B_{p^*}$. Denote $L_{(\pi,p^*)} = \nabla_{(\pi',p^{*\prime})} L(\mathbf{P}^0, \pi^0, p^{*0})$. Then, the convergence properties of the NPL algorithm corresponding to those of Lemma 2 is given by: for $j = 1, \ldots, k$,

$$\begin{aligned} \tilde{\zeta}_j^* - \hat{\zeta}_{NPL}^* &= O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||), \\ \tilde{\mathbf{P}}_j - \hat{\mathbf{P}}_{NPL} &= [I - \mathbf{\Psi}_\theta D_{(\pi,p^*)} \mathbf{\Psi}_\theta' L_P' \Delta_L^{1/2} M_{L_{(\pi,p^*)}} \Delta_L^{1/2} L_P] \mathbf{\Psi}_P (\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) \\ &\quad + O_p(n^{-1/2} ||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||) + O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||^2). \end{aligned}$$

where $D_{(\pi,p^*)} = (\mathbf{\Psi}_\theta' L_P' \Delta_L^{1/2} M_{L_{(\pi,p^*)}} \Delta_L^{1/2} L_P \mathbf{\Psi}_\theta)^{-1}$; $M_{L_{(\pi,p^*)}} = I - \Delta_L^{1/2} L_{(\pi,p^*)} (L_{(\pi,p^*)}' \Delta_L L_{(\pi,p^*)})^{-1} L_{(\pi,p^*)} \Delta_L^{1/2}$. The proof is similar to that of Lemma 2 and omitted.

6.2 Time-dependent unobserved heterogeneity and the EM algorithm

We now extend our analysis to models with time-varying unobserved states using the approach developed by Arcidiacono and Miller (2008). The setup is similar to that of the previous section except that unobserved states transition over time.

Suppose that there are observed state variables, $x_{it} \in X = \{1, ..., |X|\}$, and unobserved state variables, $s_{it} \in S = \{1, ..., M\}$. We assume that unobserved state variable s_{it} follows an exogenous first-order Markov process where $\pi(j,k)$ is the transition probability from state j to k. Let Π denote the $M \times M$ matrix representing the transition probabilities for the unobserved states. To simplify our analysis, we assume that the transition probabilities of x, denoted by $f_x(x_{i,t+1}|a_{it}, x_{it})$, is independent of unobserved states and known to econometricians. Let $\zeta = (\theta', \text{vec}(\Pi)')' \in \Theta_\zeta$ be the parameter vector to be estimated and ζ^0 denote the true parameter.

We observe a panel data set $\{w_i\}_{i=1}^{n}$ where $w_i = \{a_{it}, x_{it}, x_{i,t+1}\}_{t=1}^{T}$ is randomly drawn across i's. The conditional probability distribution of a_{it} given x_{it} *and* s_{it} is given by a fixed point of $P_\theta = \Psi(P_\theta, \theta)$. The initial joint distribution of (x_{i1}, s_{i1}) is assumed to be randomly drawn from the stationary distribution satisfying the following equation: $p^*(x,s) = \sum_{x'=1}^{|X|} \sum_{s'=1}^{S} p^*(x', s') \left(\sum_{a'=1}^{A} P_\theta(a'|x', s') f_x(x|a', x') \pi_{s',s} \right)$. As in the previous section, we assume that the fixed point of this functional equation is available; denote the stationary distribution of (x,s) under P_θ and Π by $p_{P_\theta,\Pi}^*$. As stated in Remark 9, the stationarity assumption can be relaxed.

Given the conditional choice probabilities P and the transition matrix for the unobserved states Π, the likelihood contribution from the i-th observation is given by:

$$[L(P,\Pi)](w_i) \equiv \sum_{s_1=1}^{S}\sum_{s_2=1}^{S} \ldots \sum_{s_T=1}^{S} p^*_{P,\Pi}(x_{i1},s_1)\prod_{t=2}^{T}\pi(s_{t-1},s_t)P(a_{it}|x_{it},s_t)f_x(x_{i,t+1}|a_{it},x_{it}).$$

The NPL algorithm for models with time-dependent unobserved heterogeneity is described as follows. Let $\tilde{P}_0$ be an initial guess for the conditional choice probabilities. For $j=1,2,...$, iterate

Step 1: Given $\tilde{P}_{j-1}$, update $\zeta=(\theta',\text{vec}(\Pi)')'$ by $\tilde{\zeta}_j=\arg\max_{\zeta\in\Theta_\zeta}\ n^{-1}\sum_{i=1}^{n}\ln\Big([L(\Psi(\tilde{P}_{j-1},\theta),\Pi)](w_i)\Big)$,

Step 2: Update P using the obtained estimate $\tilde{\theta}_j$ by $\tilde{P}_j=\Psi(\tilde{P}_{j-1},\tilde{\theta}_j)$,

until $j=k$. This algorithm achieves the same convergence rate as that of Lemma 2. In particular, the dominant eigenvalue of Ψ_P determines the convergence rate. The proof of this claim is essentially the same as that of Lemma 2 and omitted.

It is often very difficult to directly solving the optimization problem in Step 1. Following Arcidiacono and Miller (2008), we may incorporate the EM algorithm into the NPL algorithm as follows. Let $s^T=(s_1,s_2,...,s_T)'\in S^T$ be a possible sequence of unobserved state variables. Given P and Π, define $[L^*(P)](w_i,s^T)=\prod_{t=2}^{T}P(a_{it}|x_{it},s_t)f_x(x_{i,t+1}|a_{it},x_{it})$ and $[\pi^T(P,\Pi)](s^T|x_{i1})=p^*_{P,\Pi}(x_{i1},s_1)\prod_{t=2}^{T}\pi(s_{t-1},s_t)$ so that $[L(P,\Pi)](w_i)=\sum_{s'\in S^T}[\pi^T(P,\Pi)](s'|x_{i1})[L^*(P)](w_i,s')$.

Given an initial guess $(\tilde{P}_0,\tilde{\zeta}_0)$, for $j=1,2,...$, iterate

Step 1: Let $\bar{\theta}_0=\tilde{\theta}_{j-1}$ and $\bar{\Pi}_0=\tilde{\Pi}_{j-1}$. Given $\tilde{P}_{j-1}$, iterate the following E-step and M-step for $r=1,2,...$

(E-step) Given $\bar{\theta}_{r-1}$ and $\bar{\Pi}_{r-1}$, compute

$$\tau_{r,i}(s^T)=\frac{[\pi^T(\Psi(\tilde{P}_{j-1},\bar{\theta}_{r-1}),\bar{\Pi}_{r-1})](s^T|x_{i1})[L^*(\Psi(\tilde{P}_{j-1},\bar{\theta}_{r-1}))](w_i,s^T)}{[L(\Psi(\tilde{P}_{j-1},\bar{\theta}_{r-1}),\bar{\Pi}_{r-1})](w_i)}$$

for $i=1,...,n$.

(M-step) Maximize the expected pseudo-log-likelihood:

$$\bar{\theta}_r=\arg\max_{\theta\in\Theta}\ n^{-1}\sum_{i=1}^{n}\sum_{s^T\in S^T}\tau_{r,i}(s^T)\ln\Big([L^*(\Psi(\tilde{P}_{j-1},\theta))](w_i,s^T)\Big).$$

and, given $\bar{\theta}_r$, update Π by

$$\bar{\Pi}_r=\arg\max_{\Pi\in\Theta_\Pi}\ n^{-1}\sum_{i=1}^{n}\sum_{s^T\in S^T}\tau_{r,i}(s^T)\ln\Big([\pi^T(\Psi(\tilde{P}_{j-1},\bar{\theta}_r),\Pi)](s^T|x_{i1})\Big)$$

until $r=R$. Update θ by $\tilde{\theta}_j=\bar{\theta}_R$.

Step 2: Update P using the obtained estimate $\tilde{\theta}_j$ by $\tilde{P}_j = \Psi(\tilde{P}_{j-1}, \tilde{\theta}_j)$,

until $j = k$. By choosing R sufficiently large or, alternatively, iterating E-step and M-step in Step 1 until convergence, this NPL algorithm with the EM algorithm produces the same sequence of estimators as the original NPL algorithm.

When the evaluation of the mapping Ψ is costly, it is computationally demanding to implement this EM algorithm. Arcidiacono and Miller (2008) show that, when the model exhibits, what they call, finite time dependence, it is possible to define the fixed point mapping Ψ so that its evaluation is not so costly. They illustrate that finite dependence covers a broad class of models, including models with renewal as in Example 2. Applying the EM algorithm developed above to models with finite dependence is often feasible.

Remark 10 *Arcidiacono and Miller (2008) suggest updating the probabilities (i.e., Step 2) only after one iteration of E-step and M-step by choosing $R = 1$ above. Such an iterative sequence may not necessarily increase the likelihood at each iteration.*

Remark 11 *Replacing $L^*(\Psi(\tilde{P}_{j-1}, \theta))$ in both E-step and M-step by $L^*(\tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}))$ and $L^*(\Phi(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}))$, respectively, we may apply the q-NPL algorithm and the approximate AFXP algorithm in the context of the EM algorithm.*

7 Monte Carlo experiments

We consider the model of Example 3. The profit function of firm i operating in market m in period t is specified as $\tilde{\Pi}_i(a_{mt}, S_{mt}, a_{m,t-1}, \epsilon_{imt}; \theta) = \Pi_i(a_{imt}, a_{-i,mt}, S_{mt}, a_{m,t-1}; \theta) + \epsilon_{imt}$ with

$$\begin{aligned}
&\Pi_i(a_{imt} = 1, a_{-i,mt}, S_{mt}, a_{m,t-1}; \theta) + \epsilon_{imt}(1) = \\
&\qquad \theta_{RS} \ln S_{mt} - \theta_{RN} \ln(1 + \sum_{j \neq i} a_{jmt}) - \theta_{FC,i} - \theta_{EC}(1 - a_{im,t-1}) + \epsilon_{imt}(1)
\end{aligned}$$

while, if the firm is not operating in market m, its profit is $\Pi_i(a_{imt} = 0, a_{-i,mt}, S_{mt}, a_{m,t-1}; \theta) + \epsilon_{imt}(0) = \epsilon_{imt}(0)$. We assume that $\{\epsilon_{imt}\}$ are i.i.d. extreme value type I with zero mean and unit variance and S_{mt} follows an exogenous first-order Markov process $f_S(S_{m,t+1}|S_{mt})$. The explicit expression for the fixed mapping Ψ in this model is provided in the Appendix B.

We set the number of firms $N = 3$. The state space for the market size S_{mt} is $\{2, 6, 10\}$ and its transition probability matrix is given by

$$\begin{bmatrix} 0.8 & 0.2 & 0.0 \\ 0.2 & 0.6 & 0.2 \\ 0.0 & 0.2 & 0.8 \end{bmatrix}.$$

The discount factor is set to $\beta = 0.96$. Fixed operating costs are $\theta_{FC,1} = 1.0$, $\theta_{FC,2} = 0.9$, and $\theta_{FC,3} = 0.8$ while we set both θ_{RS} and θ_{EC} equal to 1.

The value of parameter θ_{RN} determines the degree of strategic substitutabilities among firms and is the main determinant of the dominant eigenvalues of Ψ_P. We therefore vary the values of θ_{RN} across experiments and examine the performance of different estimators across different parameter values of θ_{RN}. In particular, we consider $\theta_{RN} = 1, 2, 4$, and 6. All the eigenvalues of Ψ_P are less than 1 in absolute value for $\theta_{RN} = 1$ and 2 while the smallest eigenvalues are less than -1 for $\theta_{RN} = 4$ and 6. The second and the third column of Table 1 respectively show the largest and the smallest eigenvalues of $\nabla_{P'}\Psi$ evaluated at the fixed point. We estimate θ_{RS} and θ_{RN} while the other parameters are not estimated but fixed at the true values.

Given the equilibrium choice probabilities obtained as the fixed point of Ψ and the transition probabilities for market size S, we obtain the steady state distribution. To generate an observation in market m, we first randomly draw $x_m = \{S_{m1}, a_{1m0}, a_{2m0}, a_{3m0}\}$ from the steady-state distribution and then, conditioning on the realized value of x_m, the choice a_{im1} for $i = 1, 2, 3$ is randomly drawn from the equilibrium choice probabilities. Repeating the procedure for $m = 1, 2, ..., M$, we obtain a data set with a sample size M: $\{S_{m1}, a_{im0}, a_{im1} : i = 1, 2, 3; m = 1, 2, ..., M\}$. In our experiment, we produce 500 simulated samples, each of which contains $M = 400$ observations.

To generate the data for each experiment, we need to compute a fixed point of $\Psi(P, \theta)$. For $\theta_{RN} = 1$ and 2, the fixed point is obtained by iterating the mapping $\Psi(P, \theta)$ starting from an initial vector of choice probabilities, P_0, with all probabilities equal to 0.5. For $\theta_{RN} = 4$ and 6, the smallest eigenvalues of Ψ_P evaluated at the fixed point are smaller than negative one at -1.18 and -1.48, respectively, and hence the sequence $\{P_k\}$ obtained by iterating the mapping $\Psi(P, \theta)$ does not converge. To obtain a fixed point of $\Psi(P, \theta)$ for $\theta_{RN} = 4$ and 6, we consider an alternative mapping $[\Lambda(P, \theta)](a = 1|x) \equiv \{[\Psi(P, \theta)](a = 1|x)\}^{\alpha^*}\{P(a = 1|x)\}^{1-\alpha^*}$ with the optimal value of α^*, which has better convergence property than Ψ.

We may estimate the absolute value of the dominant eigenvalue of Ψ or Λ by simulating a sequence $P^j = \Psi(P^{j-1}, \theta)$ or $P^j = \Lambda(P^{j-1}, \theta)$ for $j = 1, ..., J$ and computing the mean of $||P^{j+1} - P^J||/||P^j - P^J||$'s, where J is the number of iterations at convergence. The first and the second panel of Table 2 shows that the convergence rate of P computed in this way is very close to the dominant eigenvalue of Ψ or Λ across different value of θ_{RN}.

This procedure can be also used to estimate the optimal value α^* when the eigenvalue of Ψ_P is difficult to compute; we may pick up the value of $\alpha \in \{0.01, 0.02, ..., 0.99, 1.00\}$ that leads to the smallest value of the mean of $||P^{j+1} - P^J||/||P^j - P^J||$'s. The seventh column of Table 1 reports such an estimate. Comparing the seventh column with the sixth column of Table 1, an estimate of α^* using this procedure approximates the true value of α^* well. The last two columns of Table 1 report the absolute value of the dominant eigenvalue of $M_{\Psi_\theta}\Psi_P$ and $M_{\Lambda_\theta}\Lambda_P$. They are similar to the corresponding eigenvalues of Ψ_P and Λ_P reported in the second to fifth columns. Thus, in view of Lemma 1, the convergence rate of the NPL algorithm is primarily determined by the dominant eigenvalue of Ψ_P or Λ_P.

We now examine the convergence property of the NPL algorithm. For each sample, we compute the convergence rate of P as the mean of $||\tilde{P}_{j+1} - \tilde{P}_k||/||\tilde{P}_j - \tilde{P}_k||$'s, where $\{\tilde{P}_j\}_{j=1}^k$ is a sequence of estimators generated by the NPL algorithm. The first row of the last panel of Table 2 reports the median value of the convergence rates of P across 500 samples using the NPL algorithm with Ψ. For $\theta_{RN} = 1$ and 2, the median convergence rate of P in NPL algorithm is close to the absolute value of the dominant eigenvalue of Ψ_P as Lemma 1 predicts. On the other hand, for $\theta_{RN} = 4$ or 6, when we compute the convergence rates using the first 50 iterations, they are more than one, suggesting that a sequence of estimators generated by the NPL algorithm with Ψ is not converging.

The second row of the last panel of Table 2 reports the convergence rate when the mapping Ψ is replaced with a transformed mapping $\Lambda = \Psi^{\alpha^*} P^{1-\alpha^*}$. Using Λ in place of Ψ improves the convergence property of the NPL algorithm; in particular, the convergence rates are now less than one for $\theta_{RN} = 4$ and 6, implying that the NPL algorithm with Λ is converging. The last three rows of Table 2 show that using the q-NPL algorithm with Λ and $q = 3$, the NR-based q-NPL algorithm with Λ and $q = 3$, or the q-AFXP algorithm with $q = 3$ instead of the q-NPL algorithm with Λ further improves the convergence rates. These results are consistent with our theoretical analysis.

Table 3 compares the bias and the mean squared errors (MSE) across different estimators. The estimators generated by the NPL algorithm with Ψ converges to the same estimate as that with Λ for $\theta_{RN} = 1$ or 2. For $\theta_{RN} = 4$ and 6, however, reflecting its non-convergence property, the estimator generated by 50 iterations of the NPL algorithm with Ψ performs worse than those with Λ; in particular, for $\theta_{RN} = 4$ and 6, the square root of the integrated MSE for the estimates of $\hat{P}$ generated by the NPL algorithm with Ψ is more than twenty times larger than those with Λ.

The third row of each panel of Table 3 shows the performance of the estimator generated by the q-NPL algorithm with Λ and $q = 3$ while the fourth row of each panel of Table 3 reports the q-AFXP algorithm. The NR-based q-NPL algorithm converged to the same limit as the q-NPL algorithm in all simulations. The estimators generated by the q-NPL algorithm as well as the q-AFXP algorithm perform better than the estimators generated by the NPL algorithm especially for $\theta_{RN} = 2, 4$, and 6, suggesting their efficiency gains over the NPL estimator.

The fifth to the ninth rows of each panel of Table 3 reports the bias and the MSE for the following five different estimators: the 2-step PML estimators with Ψ and Λ, the 3-step q-NPL estimator obtained by taking one iteration of the q-NPL algorithm with Λ starting from the 2-step PML estimator with Ψ, the 3-step NR-q-NPL estimator that is similar to the 2-step q-NPL estimator but updated with NR-step, and the 3-step q-AFXP estimator generated by taking one additional iteration of the q-AFXP algorithm from the PML. Both the 2-step estimators with Ψ and those with Λ perform substantially worse than other 3-step estimators especially for $\theta_{RN} = 2$, 4, and 6. The results suggest that the popular two-step estimators can be very imprecise when the initial choice probabilities estimate is imprecise and taking additional iterations using either the q-NPL algorithm or the AFXP algorithm may dramatically improve the efficiency property of the estimators.

8 Appendix A: Proofs

In the following, C denotes a generic positive and finite constant, and it may take different values in different places.

8.1 Proof of Proposition 1

Assumption 1 (a), (b), and (d) with $\hat{P}_0 \to_p P^0$ imply that $\overline{\psi}(\hat{P}_0, \theta)$ converges uniformly in probability in θ to $E(\ln \Psi(P^0, \theta))$ (c.f., Lemma 24.1 of Gourieroux and Monfort, 1989). Then, the rest of the proof follows the proof of Theorem 2.1 of Newey and McFadden (1994). □

8.2 Proof of Propositions 2 and 3

See pp.49-52 of Aguirregabiria and Mira (2007). □

8.3 Proof of Lemma 1

Define $\overline{\psi}_\theta(P, \theta) \equiv n^{-1} \sum_{i=1}^n (\partial/\partial\theta) \ln \Psi(P, \theta)(a_i|x_i)$, $\overline{\psi}_{\theta P}(P, \theta) \equiv n^{-1} \sum_{i=1}^n (\partial^2/\partial\theta\partial P') \ln \Psi(P, \theta)(a_i|x_i)$, and $\overline{\psi}_{\theta\theta}(P, \theta) \equiv n^{-1} \sum_{i=1}^n (\partial^2/\partial\theta\partial\theta') \ln \Psi(P, \theta)(a_i|x_i)$.

Recall that $\tilde{\theta}_j$ satisfies the first order condition

$$\overline{\psi}_\theta(\tilde{P}_{j-1}, \tilde{\theta}_j) = 0. \tag{19}$$

Expanding this around $(\hat{P}_{NPL}, \hat{\theta}_{NPL})$ and using $\overline{\psi}_\theta(\hat{P}_{NPL}, \hat{\theta}_{NPL}) = 0$ gives

$$0 = \overline{\psi}_{\theta P}(\bar{P}, \bar{\theta})(\tilde{P}_{j-1} - \hat{P}_{NPL}) + \overline{\psi}_{\theta\theta}(\bar{P}, \bar{\theta})(\tilde{\theta}_j - \hat{\theta}_{NPL}),$$

where $(\bar{P}, \bar{\theta})$ lie between $(\tilde{P}_{j-1}, \tilde{\theta}_j)$ and $(\hat{P}_{NPL}, \hat{\theta}_{NPL})$. Inverting $\overline{\psi}_{\theta\theta}(\bar{P}, \bar{\theta})$, we obtain

$$\tilde{\theta}_j - \tilde{\theta} = -\overline{\psi}_{\theta\theta}(\bar{P}, \bar{\theta})^{-1} \overline{\psi}_{\theta P}(\bar{P}, \bar{\theta})(\tilde{P}_{j-1} - \hat{P}_{NPL}) = O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||), \tag{20}$$

where the last equality follows from the last two assumptions of Assumption 2.[10]

For the second result, expand the second-step updating equation $\tilde{P}_j = \Psi(\tilde{P}_{j-1}, \tilde{\theta}_j)$ twice around $(\hat{P}_{NPL}, \hat{\theta}_{NPL})$ and use $\Psi(\hat{P}_{NPL}, \hat{\theta}_{NPL}) = \hat{P}_{NPL}$, root-$n$ consistency of $(\hat{P}_{NPL}, \hat{\theta}_{NPL})$,

[10] If we assume $\tilde{P}_{j-1}$ is consistent, the second equality follows from consistency of $\hat{P}_{NPL}$ and $\tilde{P}_{j-1}$.

and (20), then it follows that

$$\tilde{P}_j - \hat{P}_{NPL} = \Psi_P(\tilde{P}_{j-1} - \hat{P}_{NPL}) + \Psi_\theta(\tilde{\theta}_j - \hat{\theta}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2). \tag{21}$$

Rewriting (20) by using $\overline{\psi}_{\theta P}(\hat{P}_{NPL}, \hat{\theta}_{NPL}) = -\Omega_{\theta P} + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(n^{-1/2})$ and $\overline{\psi}_{\theta\theta}(\tilde{P}_{NPL}, \hat{\theta}_{NPL}) = -\Omega_{\theta\theta} + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(n^{-1/2})$, we obtain

$$\tilde{\theta}_j - \tilde{\theta} = -\Omega_{\theta\theta}^{-1}\Omega_{\theta P}(\tilde{P}_{j-1} - \hat{P}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2).$$

Substituting this into (21) in conjunction with $\Omega_{\theta\theta}^{-1}\Omega_{\theta P} = (\Psi_\theta' \Delta_P \Psi_\theta)^{-1}\Psi_\theta' \Delta_P \Psi_P$ gives

$$\tilde{P}_j - \hat{P}_{NPL} = \left[I - \Psi_\theta(\Psi_\theta' \Delta_P \Psi_\theta)^{-1}\Psi_\theta' \Delta_P\right]\Psi_P(\tilde{P}_{j-1} - \hat{P}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}||) + O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2),$$

giving the stated result. □

8.4 Proof of Proposition 4

For any eigenvalue λ of Ψ_P, the corresponding eigenvalues of Λ_P is $\alpha\lambda + (1-\alpha) = \alpha(\lambda - 1) + 1$. We first show that, if $\lambda_{\max} > 1 > \lambda_{\min}$, then there is no value of α such that $-1 < \alpha(\lambda_{\max} - 1) + 1 < 1$ and $-1 < \alpha(\lambda_{\min} - 1) + 1 < 1$ simultaneously. Suppose $\lambda_{\max} > 1$. Then, to satisfy $-1 < \alpha(\lambda_{\max} - 1) + 1 < 1$, the value of α has to be $\frac{-2}{\lambda_{\max} - 1} < \alpha < 0$ because $\alpha(\lambda_{\max} - 1) + 1 > -1$ implies $\alpha > \frac{-2}{\lambda_{\max} - 1}$ and $\alpha(\lambda_{\max} - 1) + 1 < 1$ implies $\alpha < 0$. Similarly, if $\lambda_{\min} < 1$, then the value of α has to be $0 < \alpha < \frac{2}{1 - \lambda_{\min}}$ to satisfy $-1 < \alpha(\lambda_{\min} - 1) + 1 < 1$. Since there is no value of α satisfying $\frac{-2}{\lambda_{\max} - 1} < \alpha < 0$ and $0 < \alpha < \frac{2}{1 - \lambda_{\min}}$ simultaneously, the stated result follows.

Now, assume that $1 > \lambda_{\max} > \lambda_{\min}$. We derive the value of α that minimizes the absolute value of the dominant eigenvalue of Λ_P.

Suppose that $\alpha(\lambda_{\min} - 1) + 1 < 0$. Then, the absolute value of the dominant eigenvalue of Λ is either $\alpha(\lambda_{\max} - 1) + 1$ or $-\alpha(\lambda_{\min} - 1) - 1$. If $\alpha(\lambda_{\max} - 1) + 1 > -\alpha(\lambda_{\min} - 1) - 1$, then it is possible to reduce the value of the largest eigenvalue by increasing the value of α, and such a choice of α is not optimal. Similarly, if $\alpha(\lambda_{\max} - 1) + 1 < -\alpha(\lambda_{\min} - 1) - 1$, then such a choice of α is not optimal. Therefore, the optimal value of α satisfies $\alpha(\lambda_{\max} - 1) + 1 = -\alpha(\lambda_{\min} - 1) - 1$ and $\alpha^* = \frac{2}{2 - \lambda_{\max} - \lambda_{\min}}$. The largest and smallest eigenvalues of Λ_P with α^* is given by $\frac{\lambda_{\max} - \lambda_{\min}}{2 - \lambda_{\max} - \lambda_{\min}}$ and $-\frac{\lambda_{\max} - \lambda_{\min}}{2 - \lambda_{\max} - \lambda_{\min}}$, both of which are between -1 and 1. If $\lambda_{\max} + \lambda_{\min} > 0$, then $\lambda_{\max}$ is the dominant eigenvalue of Ψ_P and $\lambda_{\max} > \frac{\lambda_{\max} - \lambda_{\min}}{2 - \lambda_{\max} - \lambda_{\min}}$ holds. If $\lambda_{\max} + \lambda_{\min} < 0$, then $\lambda_{\min} < 0$ is the dominant eigenvalue of Ψ_P and $-\frac{\lambda_{\max} - \lambda_{\min}}{2 - \lambda_{\max} - \lambda_{\min}} > \lambda_{\min}$ holds. It follows that the absolute value of the dominant eigenvalue of Λ_P with α^* is less than that of Ψ_P.

Suppose that $\alpha(\lambda_{\min} - 1) + 1 \geq 0$. Then, the value of $\alpha(\lambda - 1) + 1 \geq 0$ for any eigenvalue λ of Ψ and $\alpha = 0$ must be the optimal choice, but this is not optimal because the value of the dominant eigenvalue of Λ_P with $\alpha = 0$ is equal to 1. □

8.5 Proof of Proposition 5

We show, for $j \geq 1$, $(\tilde{P}_j, \tilde{\theta}_j) \to_p (P^0, \theta^0)$ if $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \to_p (P^0, \theta^0)$. The stated result then follows from induction and $(\tilde{P}_0, \tilde{\theta}_0) \to_p (P^0, \theta^0)$.

Assume $(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \to_p (P^0, \theta^0)$. First, $\tilde{P}_j \to_p P^0$ follows from $\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \to_p \Lambda^q(P^0, \theta^0) = P^0$.

We proceed to show $\tilde{\theta}_j \to_p \theta^0$. Define $Q_n^q(\theta, P^*, \theta^*) = n^{-1} \sum_{i=1}^n \ln \tilde{\Lambda}^q(\theta, P^*, \theta^*)(a_i|x_i)$ and $Q^q(\theta) = E \ln \tilde{\Lambda}^q(\theta, P^0, \theta^0)(a_i|x_i)$. From Theorem 2.1 of Newey and McFadden (1994) and the compactness of Θ_j^q, the consistency of $\tilde{\theta}_j$ follows if we show

$$Q_n^q(\theta, \tilde{P}_j, \tilde{\theta}_{j-1}) - Q_n^q(\theta, P^0, \theta^0) = o_p(1) \quad \text{uniformly in } \theta \in \Theta_j^q, \tag{22}$$

$$Q_n^q(\theta, P^0, \theta^0) - Q^q(\theta) = o_p(1) \quad \text{uniformly in } \theta \in \Theta_j^q, \tag{23}$$

$$Q^q(\theta) \text{ is continuous in } \theta \text{ and uniquely maximized at } \theta^0. \tag{24}$$

We show (22) first. It follows from the mean value theorem that $Q_n^q(\theta, \tilde{P}_j, \tilde{\theta}_{j-1}) - Q_n^q(\theta, P^0, \theta^0) = D_{P,n}^q(\theta, \bar{P}, \bar{\theta})(\tilde{P}_j - P^0) + D_{\theta,n}^q(\theta, \bar{P}, \bar{\theta})(\tilde{\theta}_{j-1} - \theta^0)$, where $\bar{P} \in [\tilde{P}_j, P^0]$, $\bar{\theta} \in [\tilde{\theta}_{j-1}, \theta^0]$, and

$$D_{P,n}^q(\theta, \bar{P}, \bar{\theta}) = n^{-1} \sum_{i=1}^n \frac{\nabla_{P^{*\prime}} \tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)}{\tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)}, \quad D_{\theta,n}^q(\theta, \bar{P}, \bar{\theta}) = n^{-1} \sum_{i=1}^n \frac{\nabla_{\theta^{*\prime}} \tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)}{\tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)}.$$

Because $(\tilde{P}_j, \tilde{\theta}_{j-1}) \to_p (P^0, \theta^0)$ and $\tilde{\Lambda}^q(\theta, P^*, \theta^*)$ is continuous in (P^*, θ^*), the definition of Θ_j^q implies that, for all $(a, x) \in A \times X$,

$$\tilde{\Lambda}^q(\theta, \tilde{P}, \tilde{\theta})(a|x) \in [\varepsilon/2, 1 - \varepsilon/2] \text{ uniformly in } \tilde{P} \in [\tilde{P}_j, \theta^0], \tilde{\theta} \in [\tilde{\theta}_{j-1}, \theta^0] \text{ and } \theta \in \Theta_j^q. \tag{25}$$

Consequently, $||D_{P,n}^q(\theta, \bar{P}, \bar{\theta})|| < C||n^{-1} \sum_{i=1}^n \nabla_{P^{*\prime}} \tilde{\Lambda}^q(\theta, \bar{P}, \bar{\theta})(a_i|x_i)||$. Then, (22) follows from $(\tilde{P}_j, \tilde{\theta}_{j-1}) \to_p (P^0, \theta^0)$.

We proceed to show (23). Note that, since $\Lambda^q(P^0, \theta^0) = P^0$,

$$Q_n^q(\theta, P^0, \theta^0) = n^{-1} \sum_{i=1}^n \ln \tilde{\Lambda}^q(\theta, P^0, \theta^0)(a_i|x_i) = n^{-1} \sum_{i=1}^n \ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)(a_i|x_i).$$

Since Θ_j^q is compact and $\ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)$ is continuous in $\theta \in \Theta_j^q$, (23) follows from Lemma 2.4 of Newey and McFadden (1994) if we show $E \sup_{\theta \in \Theta_j} |\ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0) + P^0)(a_i|x_i)| < \infty$. Recall $x/(1 + x) \leq \ln(x) \leq x$ for all $x > -1$. Therefore, for all a, b such that $a + b > -1$,

$$|\ln(a + b)| = |\ln(a/b + 1)||\ln(b)| \leq \max\{|a/b|, |a/(a + b)|\}|\ln(b)|. \tag{26}$$

Apply this inequality with $a = \ln(\nabla_{\theta'} \Lambda^q(P^0, \theta^0)(\theta - \theta^0))(a_i|x_i)$ and $b = P^0(a_i|x_i)$ in conjunction with $\min_{\{a,x\}} P^0(a|x) > 0$ and (25), then it follows that

$$|\ln(\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(\theta-\theta^0)+P^0)(a_i|x_i)| \leq C|\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(\theta-\theta^0)(a_i|x_i)|.$$

Consequently, $E\sup_{\theta\in\Theta_j}|\ln(\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(\theta-\theta^0)+P^0)(a_i|x_i)| \leq CE\sup_{\theta\in\Theta_j}|\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(\theta-\theta^0)(a_i|x_i)| \leq CE||\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)||\sup_{\theta\in\Theta_j^q}||\theta-\theta^0|| < \infty$, and we complete the proof of (23).

It remains to show (24). $Q^q(\theta)$ is continuous in θ from Lemma 2.4 of Newey and McFadden (1994) and the proof of (23). Note that

$$\begin{aligned} Q^q(\theta)-Q^q(\theta^0) &= E\ln(\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(\theta-\theta^0)+P^0)(a_i|x_i) - E\ln P^0(a_i|x_i) \\ &= E\ln\left(\frac{\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0)}{P^0(a_i|x_i)}+1\right). \end{aligned}$$

We show

$$E\ln\left(\frac{\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0)}{P^0(a_i|x_i)}+1\right) < E\left[\frac{\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0)}{P^0(a_i|x_i)}\right] \quad \text{for all } \theta\neq\theta^0, \tag{27}$$

then $Q^q(\theta)-Q^q(\theta^0) < 0$ for all $\theta\neq\theta^0$ because $E[\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)/P^0(a_i|x_i)] = 0$.

Recall $\ln(1+x) \leq x$ for all $x > -1$ where the inequality is strict if $x \neq 0$. Thus, (27) holds if, for all $\theta\neq\theta^0$, we have $\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0)/P^0(a_i|x_i) \neq 0$ with positive probability. Since $P^0(a_i|x_i)$ is bounded away from both 0 and ∞, this is equivalent to: for all $\theta\neq\theta^0$, $\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(a_i|x_i)(\theta-\theta^0) \neq 0$ with positive probability. This is implied by Assumption 3. Hence, (27) holds, and (24) is shown. Therefore, $\tilde{\theta}_j \rightarrow_p \theta^0$. □

8.6 Proof of Proposition 6

To analyze $\tilde{\theta}_j$, let us introduce a simplified notation for the objective function in the jth iteration:

$$Q_n^{q(j)}(\theta) \equiv Q_n^q(\theta,\tilde{P}_j,\tilde{\theta}_{j-1}) = n^{-1}\sum_{i=1}^n \ln\tilde{\Lambda}^q(\theta,\tilde{P}_j,\tilde{\theta}_{j-1})(a_i|x_i),$$

where $\tilde{\Lambda}^q(\theta,P^*,\theta^*) = \nabla_{\theta'}\Lambda^q(P^*,\theta^*)(\theta-\theta^*)+\Lambda^q(P^*,\theta^*)$.

The estimate $\tilde{\theta}_j$ satisfies the first order condition: $\nabla_{\theta'}Q_n^{q(j)}(\tilde{\theta}_j) = 0$. Applying a second-order Taylor expansion to each element of $\nabla_{\theta'}Q_n^{q(j)}(\tilde{\theta}_j)$ around $\tilde{\theta}_{j-1}$, we obtain

$$\begin{aligned} 0 &= \nabla_{\theta'}Q_n^{q(j)}(\tilde{\theta}_j) = \nabla_{\theta'}Q_n^{q(j)}(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j-\tilde{\theta}_{j-1})'\nabla_{\theta\theta'}Q_n^{q(j)}(\tilde{\theta}_{j-1}) \\ &\quad +[(\tilde{\theta}_j-\tilde{\theta}_{j-1})'B_1(\tilde{\theta}_j-\tilde{\theta}_{j-1}),\cdots,(\tilde{\theta}_j-\tilde{\theta}_{j-1})'B_K(\tilde{\theta}_j-\tilde{\theta}_{j-1})], \end{aligned} \tag{28}$$

where B_k, $k=1,\ldots,K$, is the second derivative of the kth element of $\nabla_{\theta'}Q_n^{q(j)}(\theta)$ evaluated at

$\bar{\theta} \in [\tilde{\theta}_j, \tilde{\theta}_{j-1}]$. We find an alternate expression for the last term on the right. Note that

$$\begin{aligned}
& (\tilde{\theta}_j - \tilde{\theta}_{j-1})' B_k (\tilde{\theta}_j - \tilde{\theta}_{j-1}) \\
= \; & (\tilde{\theta}_j - \hat{\theta}_{qNPL} + \hat{\theta}_{qNPL} - \tilde{\theta}_{j-1})' B_k (\tilde{\theta}_j - \hat{\theta}_{qNPL} + \hat{\theta}_{qNPL} - \tilde{\theta}_{j-1}) \\
= \; & (\tilde{\theta}_j - \hat{\theta}_{qNPL})' B_k [(\tilde{\theta}_j - \hat{\theta}_{qNPL}) + 2(\hat{\theta}_{qNPL} - \tilde{\theta}_{j-1})] + (\hat{\theta}_{qNPL} - \tilde{\theta}_{j-1})' B_k (\hat{\theta}_{qNPL} - \tilde{\theta}_{j-1}) \\
= \; & (\tilde{\theta}_j - \hat{\theta}_{qNPL})' C_k + (\hat{\theta}_{qNPL} - \tilde{\theta}_{j-1})' B_k (\hat{\theta}_{qNPL} - \tilde{\theta}_{j-1}), \qquad (29)
\end{aligned}$$

where C_k is a $K \times K$ matrix. C_k is $o_p(1)$ for all k. Substituting this to the last term on the right of (28), we can rewrite the first order condition (28) as

$$0 = \nabla_{\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j - \tilde{\theta}_{j-1})' \nabla_{\theta\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) + (\tilde{\theta}_j - \hat{\theta}_{qNPL})' o_p(1) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2). \quad (30)$$

For the first term on the right of (30), define a $1 \times K$ vector

$$L_n^q(P, \theta) = n^{-1} \sum_{i=1}^{n} \frac{\nabla_{\theta'} \Lambda^q(P, \theta)(a_i|x_i)}{\Lambda^q(P, \theta)(a_i|x_i)},$$

then we can write $\nabla_{\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) = L_n^q(\tilde{P}_j, \tilde{\theta}_{j-1})$. Furthermore, the q-NPL estimator $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ satisfies

$$L_n^q(\hat{P}_{qNPL}, \hat{\theta}_{qNPL}) = n^{-1} \sum_{i=1}^{n} \frac{\nabla_{\theta'} \Lambda^q(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})(a_i|x_i)}{\Lambda^q(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})(a_i|x_i)} = n^{-1} \sum_{i=1}^{n} \nabla_{\theta'} \ln \Lambda^q(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})(a_i|x_i) = 0.$$

Therefore, the first term on the right of (30) is approximated as

$$\begin{aligned}
& \nabla_{\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) - 0 \\
= \; & L_n^q(\tilde{P}_j, \tilde{\theta}_{j-1}) - L_n^q(\hat{P}_{qNPL}, \hat{\theta}_{qNPL}) \\
= \; & (\tilde{P}_j - \hat{P}_{qNPL})' \nabla_P L_n^q(\hat{\theta}_{qNPL}, \hat{P}_{qNPL}) + (\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL})' \nabla_\theta L_n^q(\hat{\theta}_{qNPL}, \hat{P}_{qNPL}) \\
& + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2) + O_p(||\tilde{P}_j - \hat{P}_{qNPL}||^2) \\
= \; & (\tilde{P}_j - \hat{P}_{qNPL})' E \nabla_{P\theta'} \ln \Lambda^q(P^0, \theta^0) + (\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL})' E \nabla_{\theta\theta'} \ln \Lambda^q(P^0, \theta^0) + r_{n,j}, \quad (31)
\end{aligned}$$

where $r_{n,j}$ denotes a generic reminder term of the form

$$r_{n,j} = O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2) + O_p(n^{-1/2} ||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||) + O_p(||\tilde{P}_j - \hat{P}_{qNPL}||^2) + O_p(n^{-1/2} ||\tilde{P}_j - \hat{P}_{qNPL}||),$$

and the last equality follows from expanding $\nabla_P L_n(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ and $\nabla_\theta L_n(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ around (P^0, θ^0) and using the root-n consistency of $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$.

For the second term on the right of (30), define a $1 \times K$ vector $g_i^q = \nabla_{\theta'} \Lambda^q(\tilde{P}_j, \tilde{\theta}_{j-1})(a_i|x_i)$, then

$$\nabla_{\theta\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) = -n^{-1} \sum_{i=1}^{n} \frac{g_i^{q\prime} g_i^q}{(\Lambda^q(\tilde{P}_j, \tilde{\theta}_{j-1})(a_i|x_i))^2}.$$

Therefore, in view of the root-n consistency of $\hat{\theta}_{qNPL}$, we obtain

$$\begin{aligned} \nabla_{\theta\theta'} Q_n^{q(j)}(\tilde{\theta}_{j-1}) &= -E[\nabla_\theta \ln \Lambda^q(P^0, \theta^0)(a_i|x_i) \nabla_{\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] \\ &\quad + O_p(n^{-1/2}) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||) + O_p(||\tilde{P}_j - \hat{P}_{qNPL}||). \end{aligned} \tag{32}$$

Substituting (31) and (32) into (30) and using $E[\nabla_\theta \ln \Lambda^q(P^0, \theta^0)(a_i|x_i) \nabla_{\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] + E[\nabla_{\theta\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] = 0$ gives

$$\begin{aligned} & \{E[\nabla_\theta \ln \Lambda^q(P^0, \theta^0)(a_i|x_i) \nabla_{\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] + O_p(n^{-1/2})\}(\tilde{\theta}_j - \hat{\theta}_{qNPL}) \\ = \;& E[\nabla_{\theta P'} \Lambda^q(P^0, \theta^0)(a_i|x_i)](\tilde{P}_j - \hat{P}_{qNPL}) + r_{n,j}. \end{aligned} \tag{33}$$

It follows that $\tilde{\theta}_j - \hat{\theta}_{qNPL} = O_p(||\tilde{P}_j - \hat{P}_{qNPL}||)$.

To obtain the updating formula of $\tilde{P}_j$, expand $\Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ around $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ and use the root-n consistency of $(\hat{P}_{qNPL}, \hat{\theta}_{qNPL})$ to get

$$\tilde{P}_j = \Lambda^q(\tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = \hat{P}_{qNPL} + \Lambda_P^q(\tilde{P}_{j-1} - \hat{P}_{qNPL}) + \Lambda_\theta^q(\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}) + r_{n,j}, \tag{34}$$

where $\Lambda_P^q \equiv \nabla_{P'} \Lambda^q(P^0, \theta^0)$ and $\Lambda_\theta^q \equiv \nabla_{\theta'} \Lambda^q(P^0, \theta^0)$.

Using matrix notations of $E[\nabla_\theta \ln \Lambda^q(P^0, \theta^0)(a_i|x_i) \nabla_{\theta'} \ln \Lambda^q(P^0, \theta^0)(a_i|x_i)] = (\Lambda_\theta^q)' \Delta_P \Lambda_\theta^q$ and $E[\nabla_{\theta P'} \Lambda^q(P^0, \theta^0)(a_i|x_i)] = -(\Lambda_\theta^q)' \Delta_P \Lambda_P^q$, (33) is written as $\tilde{\theta}_j - \hat{\theta}_{qNPL} = -\{(\Lambda_\theta^q)' \Delta_P \Lambda_\theta^q + O_p(n^{-1/2})\}^{-1} (\Lambda_\theta^q)' \Delta_P \Lambda_P^q (\tilde{P}_j - \hat{P}_{qNPL}) + r_{n,j}$. Substituting this expression for $\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}$ into (34) gives the stated convergence rate of $\tilde{P}_j$. □

8.7 Proof of Proposition 7

The proof is similar to the proof of Proposition 5 and omitted. □

8.8 Proof of Proposition 8

The proof is similar to the proof of Proposition 6 except that $\Lambda^q(P, \theta)$, $\nabla_{\theta'} \Lambda^q(P, \theta)$, and $\nabla_{P'} \Lambda^q(P, \theta)$ are replaced with P_θ, $\nabla_{\theta'} P_\theta$ and $\nabla_{P'} P_\theta = 0$, respectively, and omitted. Note that $\nabla_{P'} P_{\theta^0} = 0$ in this proposition corresponds to $\Lambda_P^q = 0$ in Proposition 6.

8.9 Proof of Proposition 9

The proof is similar to that of Proposition 5 and omitted. □

8.10 Proof of Proposition 10

The updating formula of $\tilde{P}_j$ follows simply from expanding $\Lambda^q(\tilde{P}_{j-1},\tilde{\theta}_{j-1})$ around $(\hat{P}_{MLE},\hat{\theta}_{MLE})$:

$$\begin{aligned}\tilde{P}_j &= \Lambda^q(\tilde{P}_{j-1},\tilde{\theta}_{j-1})\\ &= \hat{P}_{MLE}+\nabla_{P'}\Lambda^q(P^0,\theta^0)(\tilde{P}_{j-1}-\hat{P}_{MLE})+\nabla_{\theta'}\Lambda^q(P^0,\theta^0)(\tilde{\theta}_{j-1}-\hat{\theta}_{MLE})+O_p(||\tilde{P}_{j-1}-\hat{P}_{MLE}||^2)\\ &\quad +O_p(||\tilde{\theta}_{j-1}-\hat{\theta}_{MLE}||^2)+O_p(n^{-1/2}||\tilde{P}_{j-1}-\hat{P}_{MLE}||)+O_p(n^{-1/2}||\tilde{\theta}_{j-1}-\hat{\theta}_{MLE}||),\end{aligned}$$

where the order of $O_p(\cdot)$ terms in the second equality follows from Assumption 7(b) and the root-n consistency of $(\hat{P}_{MLE},\hat{\theta}_{MLE})$.

Define the objective function in the jth iteration by

$$Q_n^{(j)}(\theta)\equiv Q_n(\theta,\tilde{P}_j,\tilde{\theta}_{j-1})=n^{-1}\sum_{i=1}^n \ln\Phi(\theta,\tilde{P}_j,\tilde{\theta}_{j-1})(a_i|x_i),$$

where $\Phi(\theta,P^*,\theta^*)=(I-\nabla_{P'}\Psi(P^*,\theta^*))^{-1}\nabla_{\theta'}\Psi(P^*,\theta^*)(\theta-\theta^*)+P^*$ as defined in (14). The estimator $\tilde{\theta}_j$ satisfies the first order condition: $\nabla_{\theta'}Q_n^{(j)}(\tilde{\theta}_j)=0$. Apply a second-order Taylor expansion to each element of $\nabla_{\theta'}Q_n^{(j)}(\tilde{\theta}_j)$ around $\tilde{\theta}_{j-1}$, then we obtain

$$\begin{aligned}0 &= \nabla_{\theta'}Q_n^{(j)}(\tilde{\theta}_j)=\nabla_{\theta'}Q_n^{(j)}(\tilde{\theta}_{j-1})+(\tilde{\theta}_j-\tilde{\theta}_{j-1})'\nabla_{\theta\theta'}Q_n^{(j)}(\tilde{\theta}_{j-1})\\ &\quad +[(\tilde{\theta}_j-\tilde{\theta}_{j-1})'B_1(\tilde{\theta}_j-\tilde{\theta}_{j-1}),\cdots,(\tilde{\theta}_j-\tilde{\theta}_{j-1})'B_K(\tilde{\theta}_j-\tilde{\theta}_{j-1})],\end{aligned}\tag{35}$$

where B_k, $k=1,\dots,K$, is the second derivative of the kth element of $\nabla_{\theta'}Q_n^{(j)}(\theta)$ evaluated at $\bar{\theta}\in[\tilde{\theta}_j,\tilde{\theta}_{j-1}]$. We find an alternate expression for the last term on the right. Note that

$$\begin{aligned}&(\tilde{\theta}_j-\tilde{\theta}_{j-1})'B_k(\tilde{\theta}_j-\tilde{\theta}_{j-1})\\ =\ &(\tilde{\theta}_j-\hat{\theta}_{MLE}+\hat{\theta}_{MLE}-\tilde{\theta}_{j-1})'B_k(\tilde{\theta}_j-\hat{\theta}_{MLE}+\hat{\theta}_{MLE}-\tilde{\theta}_{j-1})\\ =\ &(\tilde{\theta}_j-\hat{\theta}_{MLE})'B_k[(\tilde{\theta}_j-\hat{\theta}_{MLE})+2(\hat{\theta}_{MLE}-\tilde{\theta}_{j-1})]+(\hat{\theta}_{MLE}-\tilde{\theta}_{j-1})'B_k(\hat{\theta}_{MLE}-\tilde{\theta}_{j-1})\\ =\ &(\tilde{\theta}_j-\hat{\theta}_{MLE})'C_k+(\hat{\theta}_{MLE}-\tilde{\theta}_{j-1})'B_k(\hat{\theta}_{MLE}-\tilde{\theta}_{j-1}),\end{aligned}\tag{36}$$

where C_k is a $K\times K$ matrix. C_k is $o_p(1)$ for all k because $\tilde{\theta}_j$, $\tilde{\theta}_{j-1}$, and $\hat{\theta}$, are consistent and Assumption 7(c) implies $B_k=O_p(1)$ for all k. Substituting this to the last term on the right of (35), we can rewrite the first order condition (35) as

$$0=\nabla_{\theta'}Q_n^{(j)}(\tilde{\theta}_{j-1})+(\tilde{\theta}_j-\tilde{\theta}_{j-1})'\nabla_{\theta\theta'}Q_n^{(j)}(\tilde{\theta}_{j-1})+(\tilde{\theta}_j-\hat{\theta}_{MLE})'o_p(1)+O_p(||\tilde{\theta}_{j-1}-\hat{\theta}_{MLE}||^2).\tag{37}$$

For the first term on the right of (37), define a $1\times K$ vector

$$L_n(P,\theta)=n^{-1}\sum_{i=1}^n\frac{[(I-\nabla_{P'}\Psi(P,\theta))^{-1}\nabla_{\theta'}\Psi(P,\theta)](a_i|x_i)}{P(a_i|x_i)},$$

then we can write $\nabla_{\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) = L_n(\tilde{P}_j, \tilde{\theta}_{j-1})$. Furthermore, the MLE $(\hat{P}_{MLE}, \hat{\theta}_{MLE})$ satisfies

$$\begin{aligned} L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE}) &= n^{-1} \sum_{i=1}^n \frac{[(I - \nabla_{P'}\Psi(\hat{P}_{MLE}, \hat{\theta}_{MLE}))^{-1} \nabla_{\theta'}\Psi(\hat{P}_{MLE}, \hat{\theta}_{MLE})](a_i|x_i)}{\hat{P}_{MLE}(a_i|x_i)} \\ &= n^{-1} \sum_{i=1}^n \nabla_{\theta'} \ln P_{\hat{\theta}}(a_i|x_i) = 0. \end{aligned}$$

Therefore, the first term on the right of (37) is approximated as

$$\begin{aligned} \nabla_{\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) - 0 &= L_n(\tilde{P}_j, \tilde{\theta}_{j-1}) - L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE}) \\ &= (\tilde{P}_j - \hat{P}_{MLE})' \nabla_P L_n(\hat{\theta}_{MLE}, \hat{P}_{MLE}) + (\tilde{\theta}_{j-1} - \hat{\theta}_{MLE})' \nabla_\theta L_n(\hat{\theta}_{MLE}, \hat{P}_{MLE}) \\ &\quad + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) + O_p(||\tilde{P}_j - \hat{P}_{MLE}||^2). \end{aligned} \tag{38}$$

where the order of the $O_p(\cdot)$ term follows from Assumption 7(c).

We proceed to obtain approximations of $\nabla_P L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE})$ and $\nabla_\theta L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE})$ in (38). First, $\nabla_P L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE}) = J' + O_p(n^{-1/2})$ from expanding it around (P^0, θ^0) and using the root-M consistency of $(\hat{P}_{MLE}, \hat{\theta}_{MLE})$. For $\nabla_{\theta'} L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE})$, note that

$$\begin{aligned} \nabla_\theta [\nabla_{\theta'} \ln P_\theta(a|x)] &= \nabla_\theta \left\{ \frac{[(I - \nabla_{P'}\Psi(P, \theta))^{-1} \nabla_{\theta'}\Psi(P, \theta)](a|x)}{P(a|x)} \right\} \\ &\quad + \nabla_\theta (P_\theta)' \nabla_P \left\{ \frac{[(I - \nabla_{P'}\Psi(P, \theta))^{-1} \nabla_{\theta'}\Psi(P, \theta)](a|x)}{P(a|x)} \right\}. \end{aligned}$$

Consequently, in light of $\hat{P}_{MLE} = P_{\hat{\theta}_{MLE}}$, we have $\nabla_\theta L_n(\hat{P}_{MLE}, \hat{\theta}_{MLE}) = n^{-1} \sum_{i=1}^n \nabla_{\theta\theta'} \ln P_{\hat{\theta}_{MLE}}(a_i|x_i) - \nabla_\theta (P_{\hat{\theta}_{MLE}})' \nabla_P L_n(P_{\hat{\theta}_{MLE}}, \hat{\theta}_{MLE}) = E\nabla_{\theta\theta'} \ln P_{\theta^0}(a_i|x_i) - \nabla_\theta (P_{\theta^0})' J' + O_p(n^{-1/2})$. Substituting these into the right hand side of (38) gives

$$\nabla_{\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) = (\tilde{P}_j - \hat{P}_{MLE})' J' + (\tilde{\theta}_{j-1} - \hat{\theta}_{MLE})' (E\nabla_{\theta\theta'} \ln P_{\theta^0}(a_i|x_i) - \nabla_\theta (P_{\theta^0})' J') + r_n, \tag{39}$$

where $r_n = O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) + O_p(||\tilde{P}_j - \hat{P}_{MLE}||^2) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||) + O_p(n^{-1/2}||\tilde{P}_j - \hat{P}_{MLE}||)$.

For the second term on the right of (37), define a $1 \times K$ vector $g_i = [(I - \nabla_{P'}\Psi(\tilde{P}_j, \tilde{\theta}_{j-1}))^{-1} \nabla_{\theta'}\Psi(\tilde{P}_j, \tilde{\theta}_{j-1})](a_i|x_i)$, then $\nabla_{\theta\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) = -n^{-1} \sum_{i=1}^n \frac{g_i' g_i}{(\tilde{P}_j(a_i|x_i))^2}$. Therefore, in view of Assumption 7(c), $\hat{P}_{MLE} = P_{\hat{\theta}_{MLE}}$, and the root-$n$ consistency of $\hat{\theta}_{MLE}$, we obtain

$$\begin{aligned} \nabla_{\theta\theta'} Q_n^{(j)}(\tilde{\theta}_{j-1}) &= -E[\nabla_\theta \ln P_{\theta^0}(a_i|x_i) \nabla_{\theta'} \ln P_{\theta^0}(a_i|x_i)] \\ &\quad + O_p(n^{-1/2}) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||) + O_p(||\tilde{P}_j - \hat{P}_{MLE}||). \end{aligned} \tag{40}$$

Substituting (39) and (40) into (37) and using $E[\nabla_\theta \ln P_{\theta^0}(a_i|x_i) \nabla_{\theta'} \ln P_{\theta^0}(a_i|x_i)]$

$+ E[\nabla_{\theta\theta'} \ln P_{\theta^0}(a_i|x_i)] = 0$ gives

$$\begin{aligned} & \{E[\nabla_\theta \ln P_{\theta^0}(a_i|x_i)\nabla_{\theta'} \ln P_{\theta^0}(a_i|x_i)] + o_p(1)\}(\tilde{\theta}_j - \hat{\theta}_{MLE}) \\ = \; & -J\nabla_{\theta'} P_{\theta^0}(\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}) + J(\tilde{P}_j - \hat{P}_{MLE}) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||^2) + O_p(||\tilde{P}_j - \hat{P}_{MLE}||^2) \\ & + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{MLE}||) + O_p(n^{-1/2}||\tilde{P}_j - \hat{P}_{MLE}||), \end{aligned}$$

giving the stated result. □

8.11 Proof of Proposition 11

The marginal conditions are given by

$$\begin{aligned} \bar{G}_\theta(\Psi(\tilde{P},\tilde{\theta}))'\hat{W}\bar{g}(\Psi(\tilde{P},\tilde{\theta})) &= 0, \\ \tilde{P} - \Psi(\tilde{P},\tilde{\theta}) &= 0. \end{aligned}$$

Expanding $\bar{g}(\Psi(\tilde{P},\tilde{\theta}))$ around (P^0,θ^0) and using $||\hat{f}_x - f_x|| = O_p(n^{-1/2})$ give

$$\begin{aligned} G_\theta' W\bar{g}(\Psi(P^0,\theta^0)) + G_\theta' W G_\theta(\tilde{\theta} - \theta^0) + G_\theta' W G_P(\tilde{P} - P^0) &= o_p(n^{-1/2}), \\ (I - \Psi_P)(\tilde{P} - P^0) - \Psi_\theta(\tilde{\theta} - \theta^0) &= o_p(n^{-1/2}). \end{aligned}$$

Eliminating $(\tilde{P} - P^0)$ from these equations and using $G_\theta' W G_\theta + G_\theta' W G_P (I - \Psi_P)^{-1}\Psi_\theta = G_\theta' W G_\theta^\infty$, where $G_\theta^\infty = (\partial/\partial\theta')\bar{g}(P_{\theta^0}) = -\Gamma\Delta_x(I - \Psi_P)^{-1}\Psi_\theta$, we have

$$\sqrt{n}(\tilde{\theta} - \theta^0) \to_d N(0, (G_\theta' W G_\theta^\infty)^{-1} G_\theta' W \Omega W' G_\theta((G_\theta^\infty)' W' G_\theta)^{-1}),$$

where $\Omega = E[g(a_i, x_i; P^0)g(a_i, x_i; P^0)']$. □

8.12 Proof of Proposition 12

Recall that $\tilde{\theta}_j$ satisfies the first order condition

$$\bar{G}_\theta(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{g}(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j)) = 0. \tag{41}$$

Expanding $\bar{g}(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))$ around $(\tilde{P},\tilde{\theta})$ in (41) and using $\bar{G}_\theta'(\Psi(\tilde{P},\tilde{\theta}))\hat{W}\bar{g}(\Psi(\tilde{P},\tilde{\theta})) = 0$ gives

$$\begin{aligned} \tilde{\theta}_j - \tilde{\theta} &= [\bar{G}_\theta'(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{G}_\theta(\Psi(\bar{P},\bar{\theta})) + o_p(1)]^{-1}[\bar{G}_\theta'(\Psi(\tilde{P}_{j-1},\tilde{\theta}_j))\hat{W}\bar{G}_P(\Psi(\bar{P},\bar{\theta})) + o_p(1)](\tilde{P}_{j-1} - \tilde{P}) \\ &= O_p(||\tilde{P}_{j-1} - \tilde{P}||), \end{aligned} \tag{42}$$

with $(\bar{P},\bar{\theta})$ between $(\tilde{P}_{j-1},\tilde{\theta}_j)$ and $(\tilde{P},\tilde{\theta})$.

For the second result, first, using (42), we obtain the same approximation as (21):

$$\tilde{P}_j - \tilde{P} = \Psi_P(\tilde{P}_{j-1} - \tilde{P}) + \Psi_\theta(\tilde{\theta}_j - \tilde{\theta}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2) \qquad (43)$$

Expanding $\bar{g}(\Psi(\tilde{P}_{j-1}, \tilde{\theta}_j))$ in (41) twice around $(\tilde{P}, \tilde{\theta})$ and using $\bar{G}'_\theta(\Psi(\tilde{P}_{j-1}, \tilde{\theta}_j))\hat{W}\bar{g}(\Psi(\tilde{P}, \tilde{\theta})) = O_p(n^{-1/2}||\tilde{\theta}_j - \tilde{\theta}||) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||)$,

$$\bar{G}_P(\Psi(\tilde{P}, \tilde{\theta})) = G_P + O_p(n^{-1/2}), \qquad \bar{G}_\theta(\Psi(\tilde{P}, \tilde{\theta})) = G_\theta + O_p(n^{-1/2}) \qquad (44)$$

and (42) gives

$$\begin{aligned} 0 &= \bar{G}'_\theta(\Psi(\tilde{P}_{j-1}, \tilde{\theta}_j))\hat{W}G_P(\tilde{P}_{j-1} - \tilde{P}) + \bar{G}'_\theta(\Psi(\tilde{P}_{j-1}, \tilde{\theta}_j))\hat{W}G_\theta(\tilde{\theta}_j - \tilde{\theta}) \\ &\quad + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2). \qquad (45) \end{aligned}$$

Expanding $\Psi(\tilde{P}_{j-1}, \tilde{\theta}_j)$ around $(\tilde{P}, \tilde{\theta})$ and using (42) and (44) in (45), we have

$$\tilde{\theta}_j - \tilde{\theta} = -(G'_\theta \hat{W} G_\theta)^{-1} G'_\theta \hat{W} G_P(\tilde{P}_{j-1} - \tilde{P}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2),$$

Substituting this into (43) and noting that $G_\theta = -\Gamma\Delta_x\Psi_\theta$ and $G_P = -\Gamma\Delta_x\Psi_P$, we obtain

$$\tilde{P}_j - \tilde{P} = [I + \Psi_\theta(G'_\theta \hat{W} G_\theta)^{-1} G'_\theta \hat{W}\Gamma\Delta_x]\Psi_P(\tilde{P}_{j-1} - \tilde{P}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \tilde{P}||) + O_p(||\tilde{P}_{j-1} - \tilde{P}||^2),$$

and the second result follows. □

8.13 Proof of Lemma 2

The proof follows the proof of Lemma 1. Expanding the first order condition $\bar{l}_\zeta(\tilde{\mathbf{P}}_{j-1}, \tilde{\zeta}_j) = \bar{l}_\zeta(\hat{\mathbf{P}}_{NPL}, \hat{\zeta}_{NPL}) = 0$ gives

$$\tilde{\zeta}_j - \hat{\zeta}_{NPL} = -\bar{l}_{\zeta\zeta}(\bar{\mathbf{P}}, \bar{\zeta})^{-1}\bar{l}_{\zeta P}(\bar{\mathbf{P}}, \bar{\zeta})(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) = O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||). \qquad (46)$$

where $(\bar{\mathbf{P}}, \bar{\zeta})$ is between $(\tilde{\mathbf{P}}_{j-1}, \tilde{\zeta}_j)$ and $(\hat{\mathbf{P}}_{NPL}, \hat{\zeta}_{NPL})$. This gives the bound for $\tilde{\zeta}_j - \hat{\zeta}_{NPL}$. Rewriting this further using the first three assumptions of Assumption 11 gives

$$\tilde{\zeta}_j - \hat{\zeta}_{NPL} = -\Omega_{\zeta\zeta}^{-1}\Omega_{\zeta P}(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) + O_p(n^{-1/2}||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||) + O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||^2). \quad (47)$$

On the other hand, expanding the second step equation $\tilde{\mathbf{P}}_j = \mathbf{\Psi}(\tilde{\mathbf{P}}_{j-1}, \tilde{\zeta}_j)$ twice around $(\hat{\mathbf{P}}_{NPL}, \hat{\zeta}_{NPL})$, using root-n consistency of $(\hat{\mathbf{P}}_{NPL}, \hat{\zeta}_{NPL})$ and (46) give

$$\tilde{\mathbf{P}}_j - \hat{\mathbf{P}}_{NPL} = \mathbf{\Psi}_P(\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) + \mathbf{\Psi}_\zeta(\tilde{\zeta}_j - \hat{\zeta}_{NPL}) + O_p(n^{-1/2}||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||) + O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||^2), \qquad (48)$$

where $\mathbf{\Psi}_\zeta \equiv (\partial/\partial\zeta')\mathbf{\Psi}(\mathbf{P}^0, \theta^0) = [\mathbf{\Psi}_\theta, \mathbf{0}]$. Substituting (47) into (48) gives

$$\tilde{\mathbf{P}}_j - \hat{\mathbf{P}}_{NPL} = [\mathbf{\Psi}_P - \mathbf{\Psi}_\zeta \Omega_{\zeta\zeta}^{-1}\Omega_{\zeta P}](\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}) + O_p(n^{-1/2}||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||) + O_p(||\tilde{\mathbf{P}}_{j-1} - \hat{\mathbf{P}}_{NPL}||^2).$$

Note that

$$\Omega_{\zeta\zeta}^{-1} = \begin{bmatrix} D & -D\Omega_{\theta\pi}\Omega_{\pi\pi}^{-1} \\ -\Omega_{\pi\pi}^{-1}\Omega_{\pi\theta}D & \Omega_{\pi\pi}^{-1} + \Omega_{\pi\pi}^{-1}\Omega_{\pi\theta}D\Omega_{\theta\pi}\Omega_{\pi\pi}^{-1} \end{bmatrix},$$

where $D = (\mathbf{\Psi}_\theta' L_P' \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P \mathbf{\Psi}_\theta)^{-1}$ with $M_{L_\pi} = I - \Delta_L^{1/2} L_\pi (L_\pi' \Delta_L L_\pi)^{-1} L_\pi \Delta_L^{1/2}$. Then, using $\mathbf{\Psi}_\zeta = [\mathbf{\Psi}_\theta, \mathbf{0}]$ gives $\mathbf{\Psi}_\zeta \Omega_{\zeta\zeta}^{-1}\Omega_{\zeta P} = \mathbf{\Psi}_\theta D \mathbf{\Psi}_\theta' L_P' \Delta_L^{1/2} M_{L_\pi} \Delta_L^{1/2} L_P \mathbf{\Psi}_P$, and the stated result follows. □

9 Appendix B: Additional Results

9.1 Relative efficiency of NPL, q-NPL, and MLE

The variance of the NPL estimator is given by

$$\begin{aligned} V_{NPL} &= [\Omega_{\theta\theta} + \Omega_{\theta P}(I - \Psi_P)^{-1}\Psi_\theta]^{-1}\Omega_{\theta\theta}[\Omega_{\theta\theta} + \Psi_\theta(I - \Psi_P')^{-1}\Omega_{\theta P}']^{-1} \\ &= \Psi_\theta'(I - \Psi_P)^{-1}\Delta_P\Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P(I - \Psi_P')^{-1}\Psi_\theta \end{aligned}$$

while the variance of the MLE is

$$V_{MLE} = \left(E\left[\frac{\Psi_\theta'(I - \Psi_P)^{-1}(a|x)}{P_\theta(a|x)} \frac{(I - \Psi_P')^{-1}\Psi_\theta(a|x)}{P_\theta(a|x)} \right] \right)^{-1} = \left(\Psi_\theta'(I - \Psi_P)^{-1}\Delta_P(I - \Psi_P')^{-1}\Psi_\theta \right)^{-1}.$$

Define $A = \Delta_P^{1/2}\Psi_\theta$ and $D = \Delta_P^{1/2}(I - \Psi_P)^{-1}\Psi_\theta$. Then $V_{NPL}^{-1} = D'A(A'A)^{-1}A'D$, $V_{MLE}^{-1} = D'D = D'D(D'D)^{-1}D'D$, and $V_{MLE}^{-1} - V_{NPL}^{-1} = D'[I - A(A'A)^{-1}A']D = UU'$, where $U = D'[I - A(A'A)^{-1}A']$. Therefore, $V_{MLE}^{-1} - V_{NPL}^{-1}$ is positive semi-definite.

Next, consider the variance of q-NPL estimator, denoted by V_{qNPL}. First, evaluating the derivatives at $P = P_\theta$, we have $\Psi_\theta^q \equiv \nabla_{\theta'}\Psi^q(P_\theta, \theta) = (I - \Psi_P)^{-1}(I - \Psi_P^q)\Psi_\theta$ and $\Psi_P^q \equiv \nabla_{P'}\Psi^q(P_\theta, \theta) = (\Psi_P)^q$. Taking a derivative of $P_\theta = \Psi^q(P_\theta, \theta) = \Psi(P_\theta, \theta)$ with respect to θ gives $(\Psi_\theta^q)'(I - \Psi_P^q)^{-1} = \Psi_\theta'(I - \Psi_P)^{-1}$. Using this and defining $A_q \equiv \Delta_P^{1/2}\Psi_\theta^q = \Delta_P^{1/2}(I - \Psi_P)^{-1}(I - \Psi_P^q)\Psi_\theta$, we have $V_{qNPL}^{-1} = D'A_q(A_q'A_q)^{-1}A_q'D$. It follows that $V_{MLE}^{-1} - V_{qNPL}^{-1} = U_qU_q'$ with $U_q = D'[I - A_q(A_q'A_q)^{-1}A_q']$.

Note that $D - A_q = \Delta_P^{1/2}(I - \Psi_P)^{-1}\Psi_P^q\Psi_\theta = O(|\lambda^*|^q)$, where λ^* is the dominant eigenvalue of Ψ_P. If all the eigenvalues of Ψ_P are less than one in absolute value, then $A_q \to D$ as $q \to \infty$ so that $V_{qNPL} \to V_{MLE}$ as $q \to \infty$. Expanding $D'A_q(A_q'A_q)^{-1}A_q'D$ around $A_q = D$ gives $V_{qNPL}^{-1} - V_{MLE}^{-1} = O(||A_q - D||^2) = O(|\lambda^*|^{2q})$.

9.2 The first order condition of (9) with Ψ and Λ

Without loss of generality, let $A = \{1, 2, ..., J\}$. Then, using that $[\Psi(P, \theta)](J|x) = 1 - \sum_{j=1}^{J-1}[\Psi(P, \theta)](j|x)$, the first order condition of the maximization problem in (9) is given by

$$n^{-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{J-1}\frac{1(a_i = j)[\nabla_{\theta'}\Psi(P, \theta)](j|x_i)}{[\Psi(P, \theta)](j|x_i)} - \frac{1(a_i = J)\sum_{s=1}^{J-1}[\nabla_{\theta'}\Psi(P, \theta)](s|x_i)}{1 - \sum_{s=1}^{J-1}[\Psi(P, \theta)](s|x)}\right) = 0.$$

When the mapping Ψ is replaced with $\Lambda(P, \theta) = \{\Psi(P, \theta)\}^{\alpha}P^{1-\alpha}$, the corresponding first order condition becomes $n^{-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{J-1}\frac{1(a_i=j)[\nabla_{\theta'}\Lambda(P,\theta)](j|x_i)}{[\Lambda(P,\theta)](j|x_i)} - \frac{1(a_i=J)\sum_{s=1}^{J-1}[\nabla_{\theta'}\Lambda(P,\theta)](s|x_i)}{1-\sum_{s=1}^{J-1}[\Lambda(P,\theta)](s|x)}\right) = 0$, where $\nabla_{\theta'}\Lambda(P, \theta) = \alpha\{\Psi(P, \theta)\}^{\alpha-1}P^{1-\alpha}\nabla_{\theta'}\Psi(P, \theta)$. Evaluated at the fixed point $\hat{P}_{NPL} = \Psi(\hat{P}_{NPL}, \hat{\theta}_{NPL}) = \Lambda(\hat{P}_{NPL}, \hat{\theta}_{NPL})$, we have $\nabla_{\theta'}\Lambda(\hat{P}_{NPL}, \hat{\theta}_{NPL}) = \alpha\nabla_{\theta'}\Psi(\hat{P}_{NPL}, \hat{\theta}_{NPL})$ and these two first order conditions becomes identical.

9.3 Fixed point mapping Ψ for Monte Carlo Experiments

Denote equilibrium best response probabilities by $P^* = \{P_i^*(a_i|x), i = 1, ..., N\}$ and firm's value functions associated with this equilibrium by $V_1^{P^*}, ..., V_N^{P^*}$. Then,

$$V_i^{P^*}(x_t) = \sum_{a_{it}\in A} P_i^*(a_{it}|x_t)[\pi_i^{P^*}(a_{it}, x_t; \theta) + e_i^{P^*}(a_{it}, x_t)] + \beta \sum_{x_{t+1}\in X} V_i^{P^*}(x_{t+1})f^{P^*}(x_{t+1}|x_t)$$

where $e_i^{P^*}(a_{it}, x_t) = \text{Euler's constant} - \ln(P_i^*(a_{it}, x_t))$, $\pi_i^{P^*}(a_{it}, x_t; \theta) = \sum_{a_{-i}\in A^{N-1}}\left(\prod_{j\neq i} P_j^*(a_j|x_t)\right)$ $\Pi(a_{it}, a_{-i}, x_t; \theta)$, and $f^{P^*}(x_{t+1}|x_t) = \left(\prod_{j=1}^{N} P_j^*(a_{jt}|x_t)\right) f_S(S_{t+1}|S_t)$.

We now derive the fixed point mapping Ψ for this model. In terms of matrix notation, denote $F_S = \{f_S(S'|S)\}$, $P_i = \{P_i(a|x)\}$, $P_{-i} = \{\prod_{j\neq i} P_j(a_j|x)\}$, $P = \{\prod_{i=1}^{N} P_i(a_i|x)\}$, and $\iota_k = (1, ..., 1)'$ be a $k \times 1$ vector. Both $e_i^P = \gamma - \ln(P_i)$ and $\pi_i^P(\theta)$ are $|A^N||S| \times |A|$ matrices, where the (i, j)-th element represents the value of $e_i^{P^*}(a_i, x)$ and $\pi_i^{P^*}(a_i, x; \theta)$ corresponding to a pair of the i-th state variable x and the j-th choice a.

Using these notations, we may write $\sum_{a_{it}\in A} P_i^*(a_{it}|x_t)[\pi_i^{P^*}(a_{it}, x_t; \theta) + e_i^{P^*}(a_{it}, x_t)]$ as $[\pi_i^P(\theta) + e_i^P]P_i'$ while $F^P = (\iota_{|A^N|}\iota_{|A^N|}' \otimes F_S) * (P \otimes \iota_{|S|}')$ represents the transition matrix for $x_t = (a_{t-1}, S_t)$, where $*$ represents an element-by-element multiplication. The vector of values V_i^P can be computed as $V_i^P = (I - \beta F^P)^{-1}[\pi_i^P(\theta) + e_i^P]P_i' \equiv \Gamma_i(P, \theta)$.

Then, for $i = 1, 2, ..., N$, a fixed point mapping is given by

$$[\Psi_i(P, \theta)](a_i = j|x) = \frac{\exp(\pi_i^{P^*}(j, x; \theta) + \beta\sum_{x'\in X}[\Gamma_i(P, \theta)](x')f_i^{P^*}(x'|x, j))}{\sum_{a\in A}\exp(\pi_i^{P^*}(a, x; \theta) + \beta\sum_{x'\in X}[\Gamma_i(P, \theta)](x')f_i^{P^*}(x'|x, a))},$$

where $f_i^{P^*}(x_{t+1}|x_t, a_{it}) = \left(\prod_{j\neq i} P_j^*(a_{jt}|x_t)\right) f_S(S_{t+1}|S_t)$.

References

Aguirregabiria, V. and P. Mira (2002). “Swapping the nested fixed point algorithm: a class of estimators for discrete Markov decision models.” *Econometrica* 70(4): 1519-1543.

Aguirregabiria, V. and P. Mira (2007). “Sequential estimation of dynamic discrete games.” *Econometrica* 75(1): 1-53.

Arcidiacono, P. and R. A. Miller (2008). CCP estimation of dynamic discrete choice models with unobserved heterogeneity. Mimeographed, Duke university.

Bajari, P., Benkard, C.L., and Levin, J. (2007). “Estimating dynamic models of imperfect competition.” *Econometrica* 75(5): 1331-1370.

Bajari, P., V. Chernozhukov, and H. Hong (2006). “Semiparametric estimation of a dynamic game of incomplete information.” NBER Technical Working Paper 320.

Collard-Wexler, A. (2006) Demand fluctuations and plant turnover in the Ready-Mix concrete industry. Mimeographed, NYU.

Heckman, J. (1981) “The incidental parameter problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process,” in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. Manski and D. McFadden. Cambridge: MIT Press.

Hotz, J. and R. A. Miller (1993). “Conditional choice probabilities and the estimation of dynamic models.” *Review of Economic Studies* 60: 497-529.

Kasahara, H. and K. Shimotsu (2006) “Nested pseudo-likelihood estimation and bootstrap-based inference for structural discrete Markov decision models.” Queen’s University Working Paper.

Lütkepohl, H (1996) *Handbook of Matrices.* Wiley.

Newey, W. K. and D. McFadden (1994). “Large Sample Estimation and Hypothesis Testing,” in R. F. Engle and D. L. McFadden (eds.) Handbook of Econometrics, Vol. 4, Elsevier.

Pakes, A., M. Ostrovsky, and S. Berry (2007). “Simple estimators for the parameters of discrete dynamic games (with entry/exit examples).” *RAND Journal of Economics* 38(2): 373-399.

Pesendorfer, M. and P. Schmidt-Dengler (2007). Asymptotic least squares estimators for dynamic games. Mimeographed, LSE.

Rust, J. (1987). “Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher.” *Econometrica* 55(5): 999-1033.

Zeidler, E. (1986) *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems.* New York, Springer-Verlag.

Table 1: The Largest and Smallest Eigenvalues of Ψ_P and Λ_P

	Eig(Ψ_P)		Eig(Λ_P)					
θ_{RN}	λ_{max}	λ_{min}	λ_{max}	λ_{min}	α^*	$\hat{\alpha}^*$	Eig($M_{\Psi_\theta}\Psi_P$)	Eig($M_{\Lambda_\theta}\Lambda_P$)
1	0.2104	-0.3365	0.2572	-0.2572	0.9407	0.92	0.2922	0.2555
2	0.4275	-0.6925	0.4945	-0.4945	0.8830	0.83	0.5996	0.4937
4	0.7596	-1.1839	0.8017	-0.8017	0.8250	0.80	1.1788	0.8056
6	0.8914	-1.4788	0.9161	-0.9161	0.7730	0.71	1.4775	0.9150

A pair $(\lambda_{max}, \lambda_{min})$ represents the largest and the smallest eigenvalues of Ψ_P or Λ_P, while Λ_P is defined under the optimal value of α reported in the six column. The last two columns report the largest eigenvalue in absolute value of $M_{\Psi_\theta} = I - \Psi_\theta(\Psi_\theta' \Delta_P \Psi_\theta)^{-1}\Psi_\theta' \Delta_P$ and $M_{\Lambda_\theta} = I - \Lambda_\theta(\Lambda_\theta' \Delta_P \Lambda_\theta)^{-1}\Lambda_\theta' \Delta_P$.

Table 2: Convergence Rates

	$\theta_{RN} = 1$	$\theta_{RN} = 2$	$\theta_{RN} = 4$	$\theta_{RN} = 6$
# of iterations with Ψ at convergence	11	31	diverge	diverge
The largest eigenvalue of Ψ_P in absolute value	0.3365	0.6925	1.1839	1.4789
The median convergence rate of P with Ψ	0.3244	0.7039	—	—
# of iterations with Λ at convergence	9	17	49	103
The largest eigenvalue of Λ_P in absolute value	0.2572	0.4945	0.8017	0.9161
The median convergence rate of P with Λ	0.2124	0.4922	0.7882	0.9112
The median convergence rate of P in NPL with Ψ	0.3014	0.6309	6.6153	15.639
The median convergence rate of P in NPL with Λ	0.2660	0.4564	0.7691	0.8538
The median convergence rate of P in q-NPL with Λ	0.2595	0.4649	0.6162	0.7176
The median convergence rate of P in q-NPL-NR with Λ	0.2463	0.4604	0.6325	0.7156
The median convergence rate of P in q-AFXP	0.2569	0.4858	0.6247	0.6994

The result is based on 500 simulated samples. We set $q = 3$ for the approximate q-NPL and q-AFXP algorithms. For each sample, the convergence rate of P is NPL is computed as the average of $||\hat{P}^{j+1}_{MLE} - P^J||/||\hat{P}^{j}_{MLE} - P^J||$ across $j = 1, ..., J$ where J is the number of iterations at convergence. For $\theta_{RN} = 4$ and 6, the sequence $\{\hat{P}^j_{MLE}\}$ does not converge and the median "convergence rate" of P in NPL with Ψ is computed with $J = 50$.

Table 3: Bias and MSE

		$\theta_{RN} = 1$		$\theta_{RN} = 2$		$\theta_{RN} = 4$		$\theta_{RN} = 6$	
Parameter	Estimator	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$	Bias	$\sqrt{\text{MSE}}$
$\hat{\theta}_{RS}$	NPL with Ψ	0.0010	0.2251	-0.0096	0.1445	-0.0170	0.0780	0.0036	0.0726
	NPL with Λ	0.0010	0.2251	-0.0096	0.1445	0.0003	0.0664	0.0005	0.0757
	q-NPL with Λ	-0.0005	0.2229	-0.0100	0.1351	0.0001	0.0635	0.0017	0.0710
	q-AFXP	0.0006	0.2240	-0.0202	0.1271	0.0013	0.0612	0.0030	0.0704
	2S-PML with Ψ	-0.1450	0.2509	-0.2555	0.3035	-0.1413	0.1687	-0.0703	0.1105
	2S-PML with Λ	-0.1447	0.2485	-0.2707	0.3166	-0.1164	0.1503	-0.0233	0.0918
	3S-q-NPL with Λ	-0.0400	0.2619	-0.0251	0.1967	-0.0601	0.1077	-0.0282	0.0927
	3S-NR-q-NPL with Λ	-0.0190	0.2294	0.0102	0.1788	0.0726	0.1265	-0.0125	0.0772
	3S-q-AFXP	-0.0197	0.2266	-0.0010	0.1676	-0.0172	0.0731	-0.0326	0.0847
$\hat{\theta}_{RN}$	NPL with Ψ	0.0125	0.5987	-0.0217	0.5066	-0.1756	0.3047	-0.2006	0.3812
	NPL with Λ	0.0125	0.5987	-0.0217	0.5066	0.0069	0.1570	0.0384	0.3517
	q-NPL with Λ	0.0082	0.5931	-0.0223	0.4744	0.0072	0.1479	0.0276	0.3236
	q-AFXP	0.0142	0.6058	-0.0093	0.4981	0.0075	0.1454	0.0204	0.3077
	2S-PML with Ψ	-0.3636	0.6468	-0.9117	1.0829	-0.8540	0.9931	-0.6921	0.9527
	2S-PML with Λ	-0.3807	0.6511	-1.0070	1.1660	-0.9201	1.0630	-0.9207	1.1543
	3S-q-NPL with Λ	-0.0829	0.6926	-0.1027	0.7128	-0.4611	0.6706	-0.5712	0.8806
	3S-NR-q-NPL with Λ	-0.0459	0.6078	0.0173	0.6334	0.3344	0.4758	0.1613	0.4328
	3S-q-AFXP	-0.0472	0.6014	-0.0181	0.5982	-0.1575	0.2482	-0.2507	0.5071
$\hat{P}$	NPL with Ψ	0.0000	0.0024	0.0000	0.0018	-0.0076	0.0370	-0.0115	0.0654
	NPL with Λ	0.0000	0.0024	0.0000	0.0018	-0.0002	0.0013	0.0002	0.0025
	q-NPL with Λ	0.0000	0.0024	-0.0005	0.0019	-0.0002	0.0011	0.0005	0.0030
	q-AFXP	-0.0007	0.0027	-0.0071	0.0088	0.0000	0.0010	0.0007	0.0028
	2S-PML with Ψ	-0.0007	0.0233	-0.0005	0.0278	-0.0022	0.0638	-0.0044	0.1087
	2S-PML with Λ	0.0011	0.0092	0.0029	0.0174	-0.0144	0.0424	-0.0088	0.0826
	3S-q-NPL with Λ	-0.0022	0.0056	-0.0006	0.0027	-0.0010	0.0350	-0.0001	0.0485
	3S-NR-q-NPL with Λ	0.0003	0.0029	-0.0003	0.0010	-0.0012	0.0151	-0.0048	0.0314
	3S-q-AFXP	0.0003	0.0030	-0.0003	0.0009	0.0014	0.0357	0.0002	0.0340

The result is based on 500 simulated samples. The number of observations for each sample is 400.