Sequential Estimation of Structural Models with a Fixed Point Constraint*

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

Katsumi Shimotsu
Department of Economics
Hitotsubashi University
and Queen's University
shimotsu@econ.hit-u.ac.jp

November 3, 2009

Abstract

This paper considers the estimation problem of structural models for which empirical restrictions are characterized by a fixed point constraint, such as structural dynamic discrete choice models or models of dynamic games. We analyze the conditions under which the nested pseudo-likelihood (NPL) algorithm converges to a consistent estimator and derive its convergence rate. We find that the NPL algorithm may not necessarily converge to a consistent estimator when the fixed point mapping does not have a local contraction property. To address the issue of divergence, we propose alternative sequential estimation procedures that can converge to a consistent estimator even when the NPL algorithm does not.

Keywords: contraction, dynamic games, nested pseudo likelihood, recursive projection method.
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 a fixed point constraint $P = \Psi(\theta, P)$. For example, if $P = \{P(a|x)\}$ is the conditional choice

*We are grateful to the co-editor and three anonymous referees whose comments greatly improved the paper. The authors thank Victor Aguirregabiria, David Byrne, Hide Ichimura, Kenneth Judd, Vadim Marmer, Lealand Morin, Whitney Newey, and seminar participants at the Bank of Japan, FEMES, New York Camp Econometrics, NASM, SITE, Vienna Macroeconomic Workshop, Boston University, Michigan, Montreal, Hitotsubashi, HKUST, Johns Hopkins, SETA, Tokyo, UBC, UWO, Yale, Yokohama National University, and Xiamen for helpful comments. The authors thank the SSHRC for financial support.

probabilities, and the sample data are $\{a_i, x_i\}_{i=1}^n$, then maximizing $n^{-1}\sum_{i=1}^n \ln P(a_i|x_i)$ subject to $P = \Psi(\theta, P)$ gives the Maximum Likelihood Estimator (MLE, hereafter).

The fixed point constraint $P = \Psi(\theta, P)$ summarizes the set of structural restrictions of the model that is parametrized by a finite vector $\theta \in \Theta$.[1] In principle, we may estimate the parameter θ by the Nested Fixed Point (NFXP) algorithm (Rust, 1987), which repeatedly solves all the fixed points of $P = \Psi(\theta, P)$ at each parameter value to maximize the objective function with respect to θ. The major obstacle of applying such an estimation procedure lies in the computational burden of solving the fixed point problem for a given parameter.[2]

To reduce the computational cost, 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 parameter. 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 (e.g., Bajari, Benkard, and Levin, 2007; Pakes, Ostrovsky, and Berry, 2007; Pesendorfer and Schmidt-Dengler, 2008; Bajari, Chernozhukov, Hong, and Nekipelov, 2009). These two-step estimators may suffer from substantial finite sample bias, however, when the choice probabilities are poorly estimated in the first step.

To address the limitations of two-step estimators, Aguirregabiria and Mira (2002)(2007, henceforth AM07) developed a recursive extension of the two-step method of Hotz and Miller (1993), called the *nested pseudo likelihood (NPL) algorithm*. With $P = \{P(a|x)\}$ denoting the vector of conditional choice probabilities, the NPL algorithm starts from an initial estimate $\tilde{P}_0$ and iterates the following steps until $j = k$:

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

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

The estimator $\tilde{\theta}_1$ is a version of Hotz and Miller's two-step estimator, called the *pseudo maximum likelihood (PML) estimator*. As AM07 show, it is often the case that evaluating the mapping $\Psi(\theta, P)$ for a fixed value of P across different values of θ is computationally inexpensive and implementing Step 1 of the NPL algorithm is easy. This recursive method can be applied to models with unobserved heterogeneity, and the limit of the sequence of estimators is more efficient than the two-step estimators *if it converges to a consistent fixed point*.[3]

[1] Examples of the operator $\Psi(\theta, P)$ include, among others, the policy iteration operator for a single agent dynamic programming model (e.g., Rust, 1987; Hotz and Miller, 1993; Aguirregabiria and Mira, 2002; Kasahara and Shimotsu, 2008), the best response mapping of a game (e.g., Aguirregabiria and Mira, 2007; Pakes, Ostrovsky and Berry, 2007; Pesendorfer and Schmidt-Dengler, 2008), and the fixed point operator for a recursive competitive equilibrium (e.g., Aiyagari, 1994; Krusell and Smith, 1998).

[2] Su and Judd (2008) proposed a method that does not require solving all the fixed points of $P = \Psi(\theta, P)$ at each trial value of θ.

[3] Two-step estimators can be applied to models with unobserved heterogeneity when an initial consistent estimator of the type-specific conditional choice probabilities are available. Kasahara and Shimotsu (2009) derived sufficient conditions for nonparametric identification of a finite mixture model of dynamic discrete choices.

While the NPL algorithm provides an attractive apparatus for empirical researchers, its convergence is a concern, as recognized by AM07 (p. 19). Indeed, little is known about its convergence properties except that, in some examples, the NPL algorithm converges to a point distance away from the true value as shown in Pesendorfer and Schmidt-Dengler (2008)(2009, henceforth PS09). In our simulations using the dynamic game model of AM07, we find that the NPL algorithm diverges away from a consistent estimator when the degree of strategic substitutability is high. In such cases, various two-step estimators can be used, but they may suffer from a large finite sample bias. In view of this mixed evidence and its practical importance, it is imperative that we understand the convergence properties of the NPL algorithm.

In the first of our two main contributions, this paper derives the conditions under which the NPL algorithm converges to a consistent estimator when it is started from a neighborhood of the true value. We show that a key determinant of the convergence of the NPL algorithm is the *contraction* property of the mapping Ψ. Intuitively, the faster the mapping achieves contraction, the closer the value obtained after one iteration is to the fixed point, and the NPL algorithm works well if the mapping satisfies a good contraction property.

As our second contribution, we propose alternative sequential algorithms that are implementable even when the original NPL algorithm does not converge to a consistent estimator. The first estimator replaces $\Psi(\theta, P)$ in the NPL algorithm with $\Lambda(\theta, P) = [\Psi(\theta, P)]^{\alpha} P^{1-\alpha}$, which has a better contraction property than Ψ under some conditions. The second algorithm decomposes the space of P into the unstable subspace and its orthogonal complement based on the eigenvectors of $\partial\Psi(\theta, P)/\partial P'$. It then constructs a contractive mapping by taking a Newton step on the unstable subspace. The third algorithm defines a pseudo-likelihood function in terms of multiple iterations of a fixed point mapping and, upon convergence, generates a more efficient estimator.

In the rest of the paper, Section 2 introduces a class of models with a fixed point constraint, and Section 3 analyzes the convergence properties of the NPL algorithm. Section 4 develops alternative algorithms. Simulation results are reported in Section 5, and the conclusion follows. The proofs are collected in the Appendix.

2 Maximum likelihood estimation

We consider a class of parametric discrete choice models of which restrictions are characterized by fixed point problems. Let $a_i \in A$ denote the choice variable and $x_i \in X$ the conditioning variable. Let $P(a_i|x_i)$ denote the conditional choice probability, and define $P = \{P(a|x) : (a, x) \in A \times X\}$.[4] The model is parametrized with a K-dimensional vector θ, and the fixed point constraint $P = \Psi(\theta, P)$ summarizes the restrictions of the model. For each θ, the operator

[4]The exact formulation of $P(a|x)$ depends on the specifics of the model of interest. In the dynamic game, a may represent actions of multiple players and P contains the conditional choice probabilities across all the players.

$\Psi(\theta, P)$ maps the space of conditional choice probabilities into itself. The true conditional choice probability P^0 is one of the fixed points of the operator $\Psi(\theta, P)$ evaluated at the true parameter value θ^0.

Upon estimating such models from the sample data $\{a_i, x_i\}_{i=1}^n$, researchers may consider the MLE with a fixed point constraint:

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

where $\mathcal{M}_\theta \equiv \{P \in B_P : \; P = \Psi(\theta, P)\}$ is the set of fixed points of $\Psi(\theta, P)$ given the value of $\theta \in \Theta \subset \mathbb{R}^K$. Here, B_P represents the space of conditional choice probabilities, and Θ is the parameter space.

As discussed in the introduction, if evaluating the mapping Ψ is costly, obtaining the MLE by the NFXP algorithm could be extremely computationally intensive. One of the major issues in estimating models with a fixed point constraint is to develop an estimator that is computationally simple and has good finite sample properties as an alternative to the MLE.

3 The nested pseudo likelihood (NPL) algorithm

3.1 Asymptotic properties of the NPL estimator

This section reviews the properties of the PML estimator and the NPL estimator as discussed in Aguirregabiria and Mira (2002, 2007). These are feasible alternatives to the MLE.

We assume that the support of (a_i, x_i) is finite, $A \times X = \{a^1, a^2, \ldots, a^{|A|}\} \times \{x^1, x^2, \ldots, x^{|X|}\}$.[5] Accordingly, P is represented by an L vector, where $L = |A||X|$. Given θ, the Jacobian $\nabla_{P'} \Psi(\theta, P)$ is an $L \times L$ matrix, where $\nabla_{P'} \equiv (\partial / \partial P')$. To save space, we denote the Jacobian matrices *evaluated at the true value* (θ^0, P^0) as $\Psi_P \equiv \nabla_{P'} \Psi(\theta^0, P^0)$ and $\Psi_\theta \equiv \nabla_{\theta'} \Psi(\theta^0, P^0)$. Let $||\cdot||$ denote the Euclidean norm.

We collect the assumptions employed in AM07. Define $Q_0(\theta, P) \equiv E \ln \Psi(\theta, P)(a_i|x_i)$, $\tilde{\theta}_0(P) \equiv \arg\max_{\theta \in \Theta} Q_0(\theta, P)$, and $\phi_0(P) \equiv \Psi(\tilde{\theta}_0(P), P)$. Define the set of population NPL fixed points as $\mathcal{Y}_0 \equiv \{(\theta, P) \in \Theta \times B_P : \theta = \tilde{\theta}_0(P) \text{ and } P = \phi_0(P)\}$. See AM07 for details. Denote the sth order derivative of a function f with respect to all of its parameters by $\nabla^s f$. Let $\mathcal{N}$ denote a closed neighborhood of (θ^0, P^0).

[5] It would be interesting to extend our analysis to models with continuously distributed variables. The asymptotic analysis of the NPL estimator in such models may become substantially complicated, however, because it involves functional derivatives of mappings such as $\tilde{\theta}_M(P)$. We conjecture that, under suitable regularity conditions, the NPL estimator is asymptotically normal and Lemma 1 holds if matrices such as Ψ_P and M_{Ψ_θ} are replaced with corresponding operators. A detailed analysis is beyond the scope of this paper and left for future research.

Assumption 1 *(a) The observations* $\{a_i, x_i : i = 1, \ldots, n\}$ *are independent and identically distributed, and* $dF(x) > 0$ *for any* $x \in X$*, where* $F(x)$ *is the distribution function of* x_i*. (b)* $\Psi(\theta, P)(a|x) > 0$ *for any* $(a, x) \in A \times X$ *and any* $(\theta, P) \in \Theta \times B_P$*. (c)* $\Psi(\theta, P)$ *is twice continuously differentiable. (d)* Θ *is compact and* B_P *is a compact and convex subset of* $[0, 1]^L$*. (e) There is a unique* $\theta^0 \in int(\Theta)$ *such that* $P^0 = \Psi(\theta^0, P^0)$*. (f)* (θ^0, P^0) *is an isolated population NPL fixed point. (g)* $\tilde{\theta}_0(P)$ *is a single-valued and continuous function of* P *in a neighborhood of* P^0*. (h) the operator* $\phi_0(P) - P$ *has a nonsingular Jacobian matrix at* P^0*.*

Assumption 1(b)(c) implies that $E \sup_{(\theta,P)\in\Theta\times B_P} ||\nabla^2 \ln \Psi(\theta, P)(a_i|x_i)||^r < \infty$ for any positive integer r. Assumption 1(g) corresponds to assumption (iv) in Proposition 2 of AM07.

The PML estimator is $\hat{\theta}_{PML} = \arg\max_{\theta\in\Theta} n^{-1} \sum_{i=1}^n \ln \Psi(\theta, \hat{P}_0)(a_i|x_i)$, where $\hat{P}_0$ is an initial consistent estimator of P^0. Proposition 1 of AM07 showed that the PML estimator is consistent under Assumption 1. Also, when $\hat{P}_0$ satisfies $\sqrt{n}(\hat{P}_0 - P^0) \to_d N(0, \Sigma)$, the PML estimator is asymptotically normal with asymptotic variance $V_{PML} = (\Omega_{\theta\theta})^{-1} + (\Omega_{\theta\theta})^{-1}\Omega_{\theta P}\Sigma(\Omega_{\theta P})'(\Omega_{\theta\theta})^{-1}$, with $\Omega_{\theta\theta} \equiv E[\nabla_\theta \ln \Psi(\theta^0, P^0)(a_i|x_i)\nabla_{\theta'} \ln \Psi(\theta^0, P^0)(a_i|x_i)]$, and $\Omega_{\theta P} \equiv E[\nabla_\theta \ln \Psi(\theta^0, P^0)(a_i|x_i) \times \nabla_{P'} \ln \Psi(\theta^0, P^0)(a_i|x_i)]$.

As discussed in the introduction, Aguirregabiria and Mira (2002, 2007) developed a recursive extension of the PML estimator called the NPL algorithm. Starting from an initial estimator of P^0, the NPL algorithm generates a sequence of estimators $\{\tilde{\theta}_j, \tilde{P}_j\}_{j=1}^k$, which we call the *NPL sequence*. If the NPL sequence converges, its limit satisfies the following conditions:

$$\check{\theta} = \arg\max_{\theta\in\Theta} n^{-1} \sum_{i=1}^n \ln \Psi(\theta, \check{P})(a_i|x_i) \quad \text{and} \quad \check{P} = \Psi(\check{\theta}, \check{P}). \tag{2}$$

A pair $(\check{\theta}, \check{P})$ that satisfies these two conditions in (2) is called an *NPL fixed point*. There could be multiple NPL fixed points. The *NPL estimator*, denoted by $(\hat{\theta}_{NPL}, \hat{P}_{NPL})$, is defined as the NPL fixed point with the highest value of the pseudo likelihood among all the NPL fixed points.

Proposition 2 of AM07 establishes the consistency of the NPL estimator $\hat{\theta}_{NPL}$ under Assumption 1. Thus, the NPL estimator is a consistent NPL fixed point. The NPL estimator is asymptotically normal with asymptotic variance $V_{NPL} = [\Omega_{\theta\theta} + \Omega_{\theta P}(I - \Psi_P)^{-1}\Psi_\theta]^{-1}\Omega_{\theta\theta}\{[\Omega_{\theta\theta} + \Omega_{\theta P}(I - \Psi_P)^{-1}\Psi_\theta]^{-1}\}'$. The NPL estimator does not depend on the initial estimator of P^0 and is more efficient than the PML estimator especially when the initial estimator of P^0 is imprecise.

While AM07 illustrate that the estimator obtained as a limit of the NPL sequence performs very well relative to the PML estimator in their simulation, they neither provide the conditions under which the NPL sequence converges to a consistent NPL fixed point nor analyze how fast the convergence occurs. On the other hand, PS09 present an example in which the NPL sequence converges to a NPL fixed point that is a distance away from the true value. To date, little is known about the conditions under which the NPL sequence converges to a consistent NPL fixed point, i.e., the NPL estimator.

3.2 Convergence properties of the NPL algorithm

We now analyze the conditions under which the NPL sequence locally converges to the NPL estimator. In other words, we are concerned with whether the NPL algorithm produces the NPL estimator when started from a neighborhood of the true value.

First, we state the regularity conditions. For matrix and nonnegative scalar sequences of random variables $\{X_n, n \geq 1\}$ and $\{Y_n, n \geq 1\}$, respectively, we write $X_n = O_p(Y_n)$ (or $o_p(Y_n)$) if $||X_n|| \leq CY_n$ for some (or all) $C > 0$ with probability arbitrarily close to one for sufficiently large n. When Y_n belongs to a family of random variables indexed by $\tau \in T$, we say $X_n = O_p(Y_n(\tau))$ (or $o_p(Y_n(\tau))$) *uniformly* in τ if the constant $C > 0$ can be chosen the same for every $\tau \in T$. For instance, in Lemma 1 below, we take $\tau = \tilde{P}_{j-1}$ and $Y_n(\tau) = ||\tilde{P}_{j-1} - \hat{P}_{NPL}||$. For $\epsilon > 0$, define a neighborhood of P^0 by $\mathcal{N}_P(\epsilon) = \{P : ||P - P^0|| < \epsilon\}$.

Assumption 2 *(a) Assumption 1 holds. (b) $\Psi(\theta, P)$ is three times continuously differentiable in $\mathcal{N}$. (c) $\Omega_{\theta\theta}$ is nonsingular.*

Let $P^0_{a,x}$ denote an $L \times 1$ vector whose elements are the probability mass function of (a_i, x_i) arranged conformably with $\Psi(a|x)$. Let $\Delta_P \equiv \text{diag}(P^0)^{-2}\text{diag}(P^0_{a,x})$.[6] With this notation, 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 states the local convergence rate of the NPL algorithm and is one of the main results of this paper.

Lemma 1 *Suppose Assumption 2 holds. Then, there exists $c > 0$ such that $\tilde{\theta}_j - \hat{\theta}_{NPL} = O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||)$ and $\tilde{P}_j - \hat{P}_{NPL} = M_{\Psi_\theta}\Psi_P(\tilde{P}_{j-1} - \hat{P}_{NPL}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}|| + ||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2)$ uniformly in $\tilde{P}_{j-1} \in \mathcal{N}_P(c)$, where $M_{\Psi_\theta} \equiv I - \Psi_\theta(\Psi_\theta' \Delta_P \Psi_\theta)^{-1}\Psi_\theta' \Delta_P$.*

Remark 1 *In single-agent dynamic models, the Jacobian matrix Ψ_P is zero (Aguirregabiria and Mira, 2002, Proposition 2). Consequently, $\tilde{P}_j - \hat{P}_{NPL} = O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{NPL}|| + ||\tilde{P}_{j-1} - \hat{P}_{NPL}||^2)$, which implies that the convergence rate is faster than linear and the NPL method is always stable at (θ^0, P^0). See Kasahara and Shimotsu (2008) for further details.*

Lemma 1 provides important insights into the local convergence of the NPL sequence to the NPL estimator. Define the spectral radius of A as $\rho(A) \equiv \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}$. Then $(M_{\Psi_\theta}\Psi_P)^k \to 0$ as $k \to \infty$ if and only if $\rho(M_{\Psi_\theta}\Psi_P) < 1$ (Horn and Johnson, 1985, Theorem 5.6.12).[7] Suppose $\rho(M_{\Psi_\theta}\Psi_P) < 1$ and $||\tilde{P}_0 - P^0||$ is small. Because each NPL updating of (θ, P) uses the same pseudo-likelihood function and the $O_p()$ terms are uniform in $\tilde{P}_{j-1} \in \mathcal{N}_P(c)$, we can recursively substitute for the $\tilde{P}_j$'s to show that $(\tilde{\theta}_k, \tilde{P}_k)$ converges to $(\hat{\theta}_{NPL}, \hat{P}_{NPL})$ as $k \to \infty$. The following Lemma formally states this convergence.

[6] In a multiplayer model of a dynamic game in which unobserved state variables are independent across players, such as the model of AM07, Δ_P is simplified as $\text{diag}(P^0)^{-1}\text{diag}(f_x)$, where f_x is an $L \times 1$ vector whose elements are the probability mass function of x_i arranged conformably with $P(a|x)$.

[7] $\rho(A) \leq ||A||$ holds for any matrix A and any matrix norm $||\cdot||$. Therefore, $||M_{\Psi_\theta}\Psi_P|| < 1$ is a sufficient but not necessary condition for the convergence of $(M_{\Psi_\theta}\Psi_P)^k$ to zero.

Lemma 2 *Suppose Assumption 2 holds and* $\rho(M_{\Psi_\theta}\Psi_P) < 1$. *Then, there exists* $c_2 > 0$ *such that* $\Pr(\lim_{k\to\infty} \tilde{P}_k = \hat{P}_{NPL}) \to 1$ *as* $n \to \infty$ *if* $||\tilde{P}_0 - P^0|| < c_2$.

When $\rho(M_{\Psi_\theta}\Psi_P) > 1$, an NPL updating moves some elements of $\tilde{P}_j$ further away from $\hat{P}_{NPL}$ on each iteration. Then the NPL sequence does not converge to $\hat{P}_{NPL}$ even if $\tilde{P}_0$ is very close to $\hat{P}_{NPL}$ unless $\tilde{P}_j - \hat{P}_{NPL}$ lies on a convergent hyperplane spanned by the eigenvectors corresponding to the the eigenvalues of $M_{\Psi_\theta}\Psi_P$ lying inside the unit circle. Since such a hyperplane has zero Lebesgue measure in $\mathbb{R}^L$, the probability that $\tilde{P}_{j-1} - \hat{P}$ lies on a locally convergent hyperplane approaches zero with the sample size if the limiting distribution of $n^{1/2}(\tilde{P}_{j-1} - \hat{P})$ is continuous. The case with $\rho(M_{\Psi_\theta}\Psi_P) = 1$ corresponds to a boundary case. The linear difference equation in Lemma 1 cannot fully characterize the local property of the fixed point, which depends on the details of the model (see, for example, pp. 348-351 of Strogatz (1994)).

In general, given the nonlinear nature of the mapping Ψ, its local behavior may not fully characterize its global convergence property. For instance, even when $\rho(M_{\Psi_\theta}\Psi_P) > 1$, the NPL sequence may move away from the NPL fixed point initially and then move back to the NPL fixed point or a convergent hyperplane from a distance away.[8] When the NPL sequence diverges away from the NPL estimator, an analysis of nonlinear dynamics (see, for example, Chapter 10 of Strogatz (1994)) suggests three representative possibilities. First, as PS09 illustrates, the NPL sequence may converge to a NPL fixed point that is different from the NPL estimator. Second, as our simulation suggests, it may converge to a stable cycle. Third, the NPL sequence might never settle down to a fixed point or a period orbit.

The spectral radius of $M_{\Psi_\theta}\Psi_P$ is also closely related to the consistency of the NPL estimator. Assumption 1(h), that the operator $\phi_0(P) - P$ has a nonsingular Jacobian matrix at P^0, is a key assumption for the consistency of the NPL estimator and implies Assumption 1(f).[9] The following proposition shows that $\rho(M_{\Psi_\theta}\Psi_P) < 1$ is sufficient for Assumption 1(h).

Proposition 1 *Suppose Assumption 1(a)-(e), (g) holds and* $\rho(M_{\Psi_\theta}\Psi_P) < 1$. *Then the operator* $\phi_0(P) - P$ *has a nonsingular Jacobian matrix at* P^0. *Hence, Assumption 1(f)(h) is satisfied.*

3.3 The relation between $\rho(M_{\Psi_\theta}\Psi_P)$ and $\rho(\Psi_P)$

The condition $\rho(M_{\Psi_\theta}\Psi_P) < 1$ plays an important role both for the convergence of the NPL algorithm and for the consistency of the NPL estimator. Because Ψ_P is often closely related to the characteristics of the economic model, we want to find a bound of $\rho(M_{\Psi_\theta}\Psi_P)$ in terms of $\rho(\Psi_P)$.[10] In the following, we examine the relation between $\rho(M_{\Psi_\theta}\Psi_P)$ and $\rho(\Psi_P)$.

[8] For this to occur with a nonnegligible probability, the NPL operator must map an area in $\mathbb{R}^L$ with nonzero Lebesgue measure to the NPL fixed point or a convergent hyperplane with zero Lebesgue measure. The likelihood of this occurring depends on the specifics of the model of interest.

[9] See page 21 of AM07. Our Assumption 1(f)(h) corresponds to Conditions (v)(vii) of Proposition 2 of AM07, respectively.

[10] The contraction property of Ψ may or may not be related to the stability of equilibria in the economic model. Given a model, there are often multiple ways of formulating a fixed point mapping (e.g., Hotz and Miller, 1993;

Since M_{Ψ_θ} is idempotent, M_{Ψ_θ} is diagonalizable as $M_{\Psi_\theta} = SDS^{-1}$, where the first $L-K$ diagonal elements of D are 1 and the other elements of D are zero. From the properties of the eigenvalues, we have $\rho(M_{\Psi_\theta}\Psi_P) = \rho(SDS^{-1}\Psi_P) = \rho(DS^{-1}\Psi_P S)$. In our context, typically $L \gg K$ because the dimension of the state variable is much larger than the number of parameters. Consequently, D is close to an identity matrix, and we expect that $DS^{-1}\Psi_P S \simeq S^{-1}\Psi_P S$, which implies that the dominant eigenvalues of $M_{\Psi_\theta}\Psi_P$ and Ψ_P are close to each other.[11] In our dynamic game model with $L = 144$ and $K = 2$, we find that $\rho(M_{\Psi_\theta}\Psi_P)$ is very similar to $\rho(\Psi_P)$ (see Table 1).

3.4 Simplex restriction on P

Since P represents probabilities, the elements of P must satisfy a simplex-type restriction, and this restriction needs to be imposed in parameterizing $\Psi(\theta, P)$. Consider a model with $J+1$ support points of a, then the elements of P corresponding to the $(J+1)$th element must appear in $\Psi(\theta, P)$ only implicitly as one minus the sum of the other J elements.

We may express the updating formula in Lemma 1 in terms of a smaller space by exploiting the simplex restriction as follows. Split P into P^+ and P^-, where P^+ corresponds to the first to Jth elements, whereas P^- corresponds to the $(J+1)$th element. Let $\mathbf{1}_k$ denote a k-vector of ones, then the simplex restriction implies $P^- = \mathbf{1}_{\dim(P^-)} - \mathcal{E}P^+$ for a matrix $\mathcal{E}$ of zeros and ones defined appropriately. $\Psi(\theta, P)$ satisfies an analogous simplex restriction by its construction. Split $\Psi(\theta, P)$ analogously, and write P and $\Psi(\theta, P)$ as

$$P = \begin{pmatrix} P^+ \\ P^- \end{pmatrix} = \begin{pmatrix} P^+ \\ \mathbf{1}_{\dim(P^-)} - \mathcal{E}P^+ \end{pmatrix} = P(P^+), \tag{3}$$

$$\Psi(\theta, P) = \Psi(\theta, P(P^+)) = \begin{pmatrix} \Psi^+(\theta, P^+) \\ \Psi^-(\theta, P^+) \end{pmatrix} = \begin{pmatrix} \Psi^+(\theta, P^+) \\ \mathbf{1}_{\dim(P^-)} - \mathcal{E}\Psi^+(\theta, P^+) \end{pmatrix}. \tag{4}$$

Note from (4) that the derivative of $\Psi(\theta, P)$ with respect to the $(J+1)$th element of P is zero.

As shown in the following proposition, the restrictions (3)–(4) do not affect the validity of Lemma 1, and the updating formula of P^+ completely determines the updating formula of P. Define $\Psi^+_\theta \equiv \nabla_{\theta'}\Psi^+(\theta^0, P^{0+})$ and $\Psi^+_P \equiv \nabla_{P^{+\prime}}\Psi^+(\theta^0, P^{0+})$. Define $U = [I_{\dim(P^+)} \vdots -\mathcal{E}']'$, so that $\nabla_{\theta'}\Psi(\theta, P) = U\nabla_{\theta'}\Psi^+(\theta, P^+)$, and define $\Delta^+_P \equiv U'\Delta_P U$.

Proposition 2 *Suppose $\tilde{P}_0$ satisfies the simplex restriction (3). Then Lemma 1 holds, and the updating formula for P is given by $\tilde{P}^+_j - \hat{P}^+_{NPL} = M^+_{\Psi_\theta}\Psi^+_{P^+}(\tilde{P}^+_{j-1} - \hat{P}^+_{NPL}) + O_p(n^{-1/2}||\tilde{P}^+_{j-1} - \hat{P}^+_{NPL}|| + ||\tilde{P}^+_{j-1} - \hat{P}^+_{NPL}||^2)$, where $M^+_{\Psi_\theta} = I_{\dim(P^+)} - \Psi^+_\theta(\Psi^{+\prime}_\theta\Delta^+_P\Psi^+_\theta)^{-1}\Psi^{+\prime}_\theta\Delta^+_P$, and $\tilde{P}^-_j - \hat{P}^-_{NPL} =$*

Arcidiacono and Miller, 2008) and its contraction property depends on which mapping a researcher chooses.

[11] If $\lambda(A)$ is an algebraically simple eigenvalue of A, then $\lambda(A+\Delta)/\lambda(A) = (y^H\Delta x)/(y^H Ax) + (||\Delta||^2)$, where x and y are a right- and left- $\lambda(A)$ eigenvector of A. See, for example, Theorem 6.3.12 of Horn and Johnson (1985).

$-\mathcal{E}(\tilde{P}_j^+ - \hat{P}_{NPL}^+)$. *Further,* $M_{\Psi_\theta}\Psi_P$ *and* $M_{\Psi_\theta}^+\Psi_{P^+}^+$ *have the same nonzero eigenvalues, and* Ψ_P *and* $\Psi_{P^+}^+$ *have the same nonzero eigenvalues.*

Therefore, in practice, it suffices to check the eigenvalues of $M_{\Psi_\theta}^+\Psi_{P^+}^+$ to examine the convergence property of the NPL algorithm. In the rest of the paper, we provide our theoretical results mainly in terms of Ψ and P because of its notational simplicity.

3.5 Examples

The following two examples illustrate Lemma 1 and Proposition 2.

Example 1 (A Dynamic Discrete Game by PS09) *PS09 present a game in which the global behavior of the NPL mapping can be analytically derived. We apply our local analysis to their model. We focus on* $\Psi(\theta, P)$ *and suppress the details of their model; see PS09 for details.*

In the model of PS09, there are two firms, and firm i's choice is denoted by $a_i \in \{0, 1\}$ *for* $i = 1, 2$*, where* $a_i = 1$ *indicates firm i is active. The model has no state variable, so the conditional choice probability is summarized by a two-dimensional vector* $P^+ = (P_1^+, P_2^+)'$*, where* P_i^+ *denotes firm i's probability of being active. The model has one parameter,* θ*, and the true parameter value* θ^0 *is in the interior of the parameter space* $\Theta = [-10, -1]$*.*

When P^+ *is in a neighborhood of the true value, the mapping* Ψ^+ *of this model takes the form*

$$\Psi^+(\theta, P^+) = \begin{pmatrix} \Psi_1^+(\theta, P^+) \\ \Psi_2^+(\theta, P^+) \end{pmatrix} = \begin{pmatrix} 1 + \theta P_2^+ \\ 1 + \theta P_1^+ \end{pmatrix}.$$

This model has a unique symmetric equilibrium, $P_1^+ = P_2^+ = 1/(1-\theta)$*. PS09 show there are three NPL fixed points, one of which is the NPL estimator whereas the other two NPL fixed points are inconsistent. Further, PS09 show that the NPL sequence converges to one of the inconsistent NPL fixed points if the initial estimate does not satisfy* $P_1^+ = P_2^+$*; if the initial estimate does satisfy* $P_1^+ = P_2^+$*, then the NPL sequence converges to the NPL estimator in one iteration.*

Using the framework of Lemma 1 and Proposition 2, a direct calculation gives

$$\Psi_{P^+}^+ = \begin{pmatrix} 0 & \theta^0 \\ \theta^0 & 0 \end{pmatrix}, \quad M_{\Psi_\theta}^+ = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad M_{\Psi_\theta}^+\Psi_{P^+}^+ = \frac{\theta^0}{2}\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$

The eigenvalues of $\Psi_{P^+}^+$ *are* θ^0 *and* $-\theta^0$*, and the eigenvalues of* $M_{\Psi_\theta}^+\Psi_{P^+}^+$ *are* 0 *and* $-\theta^0$*. Because all the eigenvalues of* $\Psi_{P^+}^+$ *are outside the unit circle, the fixed point mapping* $P^+ = \Psi^+(\theta, P^+)$ *has no convergent path. Multiplying* $M_{\Psi_\theta}^+$ *annihilates the eigenvector of* $\Psi_{P^+}^+$ *associated with* θ^0 *but does not change the spectral radius of* $\Psi_{P^+}^+$*. Consequently, the NPL operator inherits the instability of* $\Psi(\theta, P)$*.*

Since $\rho(M^+_{\Psi_\theta}\Psi^+_{P^+}) > 1$, there does not exist a local convergent trajectory with nonzero Lebesgue measure. The eigenvector corresponding to the zero eigenvalue is $(1,1)'$, so that the convergent trajectory is characterized by the 45 degree line $P^+_1 = P^+_2$. From Lemma 1, the NPL sequence diverges away from the NPL estimator in the neighborhood of (θ^0, P^0) if the sequence does not lie on the 45 degree line. On the other hand, if the initial estimate lies on the 45 degree line, the zero eigenvalue implies that the NPL sequence converges to the NPL estimator at a superlinear rate. These local results are weaker than the global results in PS09 but are consistent with their findings. PS09 assume $\theta^0 < -1$. But if $\theta^0 \in (-1, 0)$, then $\rho(M^+_{\Psi_\theta}\Psi^+_{P^+}) < 1$ and the NPL sequence locally converges to the NPL estimator. When $\theta^0 = -1$, then $\rho(M_{\Psi_\theta}\Psi_P) = 1$ and we cannot apply our local stability analysis.[12]

Example 2 (Stationary Distribution) *Let a_i be a random variable following a first-order Markov process, and let P denote the vector of the probability mass function of a_i. There is no conditioning variable, hence $L = |A|$ and P is $L \times 1$. Let $M(\theta)$ be the transition matrix of a_i, so that $M(\theta)$ is an $L \times L$ column stochastic matrix and each column of $M(\theta)$ belongs to a simplex. Then, the fixed point constraint for a stationarity restriction is written as $P = \Psi(\theta, P) = M(\theta)P$. As shown in Proposition 3 below, $\nabla_{P'}\Psi(\theta, P) \neq M(\theta)$ once we take into account the simplex restriction on P. Furthermore, if $M(\theta)$ is irreducible and aperiodic, all the eigenvalues of $\nabla_{P^{+\prime}}\Psi^+(\theta, P^+)$ are smaller than one in modulus. Consequently, the NPL algorithm is convergent, provided that multiplying by M_{Ψ_θ} does not change the dominant eigenvalue of Ψ_P considerably.*

Proposition 3 *Let $M(\theta)$ be an $L \times L$ column stochastic matrix, and define $\Psi(\theta, P) = M(\theta)P$. Partition $M(\theta)$ as $[M_1(\theta) \vdots M_2(\theta)]$, where $M_2(\theta)$ is $L \times 1$. Then (a) $\nabla_{P'}\Psi(\theta, P) = [M_1(\theta) - M_2(\theta)\mathbf{1}'_{L-1} \vdots 0]$, (b) the eigenvalues of $M(\theta)$ are equal to the eigenvalues of $\nabla_{P^{+\prime}}\Psi^+(\theta, P^+)$ and one, and (c) if $M(\theta^0)$ is irreducible and aperiodic, then all the eigenvalues of $\nabla_{P^{+\prime}}\Psi^+(\theta, P^+)$ are smaller than one in modulus.*

4 Alternative sequential likelihood-based estimators

When $\Psi(\theta, P)$ is not a contraction in a neighborhood of (θ^0, P^0), the NPL algorithm may not produce a consistent estimator. This section discusses alternative estimation algorithms that are implementable even in such cases.

4.1 Locally contractive mapping with the relaxation method

Consider a class of mappings that are obtained as a log-linear combination of $\Psi(\theta, P)$ and P:

$$[\Lambda(\theta, P)](a|x) \equiv \{[\Psi(\theta, P)](a|x)\}^\alpha P(a|x)^{1-\alpha}, \tag{5}$$

[12] In the model of PS09, there exists a unique globally stable population NPL fixed point when $\theta^0 = -1$.

for all $(a, x) \in A \times X$, where $\alpha \in [0, 1]$. This is called the relaxation method in numerical analysis.[13] P is a fixed point of $\Psi(\theta, P)$ if and only if it is a fixed point of $\Lambda(\theta, P)$. Further, when the real part of every eigenvalue of Ψ_P is smaller than 1, we may choose the value of α so that $\Lambda(\theta, P)$ becomes locally contractive even when $\Psi(\theta, P)$ is not locally contractive.[14] Define $\Lambda_P \equiv \nabla_{P'} \Lambda(\theta^0, P^0)$.

Proposition 4 *Suppose the real part of every eigenvalue of Ψ_P is smaller than 1. Then there exists $\alpha \in (0, 1)$ such that $\rho(\Lambda_P) < 1$.*

Consider the NPL algorithm using $\Lambda(\theta, P)$ in place of $\Psi(\theta, P)$. The advantage of this method is its computational simplicity. Since $\ln \Lambda(\theta, P) = \alpha \ln \Psi(\theta, P) + (1-\alpha) \ln P$, $n^{-1} \sum_{i=1}^n \ln \Psi(\theta, P)(a_i|x_i)$ and $n^{-1} \sum_{i=1}^n \ln \Lambda(\theta, P)(a_i|x_i)$ are maximized at the same value of θ for a given P. Thus, using Ψ and Λ gives an identical estimator and, once an appropriate value of α is determined, the NPL algorithm using Λ converges to the NPL estimator under weaker conditions than for the original NPL algorithm at the same computational cost.[15]

4.2 Recursive Projection Method

In this subsection, we construct a mapping that has a better local contraction property than Ψ, building upon the Recursive Projection Method (RPM) of Shroff and Keller (1993) (henceforth SK).

First, fix θ. Let P_θ denote an element of $M_\theta = \{P \in B_P : P = \Psi(\theta, P)\}$ so that P_θ is one of the fixed points of $\Psi(\theta, P)$ when there are multiple fixed points. Consider finding P_θ by iterating $P_j = \Psi(P_{j-1}, \theta)$ starting from a neighborhood of P_θ. If some eigenvalues of $\nabla_{P'} \Psi(\theta, P_\theta)$ are outside the unit circle, this iteration does not converge to P_θ in general. Suppose that, counting multiplicity, there are m eigenvalues of $\nabla_{P'} \Psi(\theta, P_\theta)$ that are larger than $\delta \in (0, 1)$ in modulus:

$$|\lambda_1| \geq \cdots \geq |\lambda_m| > \delta \geq |\lambda_{m+1}| \geq \cdots \geq |\lambda_L|. \tag{6}$$

Define $\mathbb{P} \subseteq \mathbb{R}^L$ as the maximum invariant subspace of $\nabla_{P'} \Psi(\theta, P_\theta)$ belonging to $\{\lambda_k\}_{k=1}^m$, and let $\mathbb{Q} \equiv \mathbb{R}^L - \mathbb{P}$ be the orthogonal complement of $\mathbb{P}$. Let Π_θ denote the orthogonal projector from $\mathbb{R}^L$ on $\mathbb{P}$. We may write $\Pi_\theta = Z_\theta Z_\theta'$, where $Z_\theta \in \mathbb{R}^{L \times m}$ is an orthonormal basis of $\mathbb{P}$. Then, for each $P \in \mathbb{R}^L$, we have the unique decomposition $P = u + v$, where $u \equiv \Pi_\theta P \in \mathbb{P}$ and $v \equiv (I - \Pi_\theta) P \in \mathbb{Q}$.

[13] Başar (1987) applies the relaxation method to find a Nash equilibrium. Ljungqvist and Sargent (2004, p. 574) also suggest applying the relaxation method to the model of Aiyagari (1994).

[14] When all the eigenvalues of Ψ_P are real, the optimal α is given by Judd (1998, p. 80) as $\alpha^* = 2/(2 - \lambda_{\max} - \lambda_{\min})$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of Ψ_P.

[15] To optimally choose the value of α, we need to evaluate the Jacobian matrix Ψ_P and all of its eigenvalues, say, using the PML estimator. In practice, when the evaluation of Ψ_P is too costly, choosing a small positive value of α leads to a locally contracting Λ from Proposition 4.

Now apply Π_θ and $I - \Pi_\theta$ to $P = \Psi(\theta, P)$, and decompose the system as follows:

$$\begin{aligned} u &= f(u, v, \theta) \equiv \Pi_\theta \Psi(\theta, u + v), \\ v &= g(u, v, \theta) \equiv (I - \Pi_\theta)\Psi(\theta, u + v). \end{aligned}$$

For a given P_{j-1}, decompose it into $u_{j-1} = \Pi_\theta P_{j-1}$ and $v_{j-1} = (I - \Pi_\theta)P_{j-1}$. Since $g(u, v, \theta)$ is contractive in v (see Lemma 2.10 of SK), we can update v_{j-1} by the recursion $v_j = g(u, v_{j-1}, \theta)$. On the other hand, when the dominant eigenvalue of Ψ_P is outside the unit circle, the recursion $u_j = f(u_{j-1}, v, \theta)$ cannot be used to update u_{j-1} because $f(u, v, \theta)$ is not a contraction in u. Instead, the RPM performs a single Newton step on the system $u = f(u, v, \theta)$, leading to the following updating procedure:

$$\begin{aligned} u_j &= u_{j-1} + (I - \Pi_\theta \nabla_{P'} \Psi(\theta, P_{j-1})\Pi_\theta)^{-1}(f(u_{j-1}, v_{j-1}, \theta) - u_{j-1}) \equiv h(u_{j-1}, v_{j-1}, \theta), \\ v_j &= g(u_{j-1}, v_{j-1}, \theta). \end{aligned} \tag{7}$$

Lemma 3.11 of SK shows that the spectral radius of the Jacobian of the stabilized iteration (7) is no larger than δ, and thus the iteration $P_j = h(\Pi_\theta P_{j-1}, (I - \Pi_\theta)P_{j-1}, \theta) + g(\Pi_\theta P_{j-1}, (I - \Pi_\theta)P_{j-1}, \theta)$ converges locally. In the following, we develop a sequential algorithm building upon the updating procedure (7).

Let $\Pi(\theta, P)$ be the orthogonal projector from $\mathbb{R}^L$ onto the maximum invariant subspace of $\nabla_{P'}\Psi(\theta, P)$ belonging to its m largest (in modulus) eigenvalues, counting multiplicity. Define u^*, v^*, $h^*(u^*, v^*, \theta)$, and $g^*(u^*, v^*, \theta)$ by replacing Π_θ in u, v, $h(u, v, \theta)$, and $g(u, v, \theta)$ with $\Pi(\theta, P)$, and define

$$\begin{aligned} \Gamma(\theta, P) &\equiv h^*(u^*, v^*, \theta) + g^*(u^*, v^*, \theta) \\ &= \Psi(\theta, P) + [(I - \Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P))^{-1} - I]\Pi(\theta, P)(\Psi(\theta, P) - P). \end{aligned} \tag{8}$$

P^0 is a fixed point of $\Gamma(\theta^0, P)$, because all the fixed points of $\Psi(\theta, P)$ are also fixed points of $\Gamma(\theta, P)$. The following proposition shows two important properties of $\Gamma(\theta, P)$: local contraction and the equivalence of fixed points of $\Gamma(\theta, P)$ and $\Psi(\theta, P)$.

Proposition 5 *(a) Suppose $I - \Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P)$ is nonsingular and hence $\Gamma(\theta, P)$ is well-defined. Then $\Gamma(\theta, P)$ and $\Psi(\theta, P)$ have the same fixed points; i.e., $\Gamma(\theta, P) = P$ if and only if $\Psi(\theta, P) = P$. (b) $\rho(\nabla_{P'}\Gamma(\theta^0, P^0)) \leq \delta^0$, where δ^0 is defined by (6) in terms of the eigenvalues of $\nabla_{P'}\Psi(\theta^0, P^0)$. Hence, $\Gamma(\theta, P)$ is locally contractive.*

Define an *RPM fixed point* as a pair $(\check{\theta}, \check{P})$ that satisfies $\check{\theta} = \arg\max_{\theta \in \Theta} n^{-1} \sum_{i=1}^n \ln \Gamma(\theta, \check{P})(a_i|x_i)$ and $\check{P} = \Gamma(\check{\theta}, \check{P})$. The *RPM estimator*, denoted by $(\hat{\theta}_{RPM}, \hat{P}_{RPM})$, is defined as the RPM fixed point with the highest value of the pseudo likelihood among all the RPM fixed points. Define the *RPM algorithm* by the same sequential algorithm as the NPL algorithm except that it uses

$\Gamma(\theta, P)$ in place of $\Psi(\theta, P)$.

Proposition 6 shows the asymptotic properties of the RPM estimator and the convergence properties of the RPM algorithm. Define the RPM counterparts of $\tilde{\theta}_0(P)$, $\phi_0(P)$, and $\Omega_{\theta\theta}$ as $\tilde{\theta}_0^\Gamma(P) \equiv \arg\max_{\theta\in\Theta} E \ln \Gamma(\theta, P)(a_i|x_i)$, $\phi_0^\Gamma(P) = \Gamma(\tilde{\theta}_0^\Gamma(P), P)$, and $\Omega_{\theta\theta}^\Gamma \equiv E\nabla_\theta \ln \Gamma(\theta^0, P^0)(a_i|x_i)\nabla_{\theta'} \ln \Gamma(\theta^0, P^0)(a_i|x_i)$. Define $\Omega_{\theta P}^\Gamma$ analogously. Define $\Gamma_P \equiv \nabla_{P'}\Gamma(\theta^0, P^0)$ and $\Gamma_\theta \equiv \nabla_{\theta'}\Gamma(\theta^0, P^0)$. We outline the assumptions first.

Assumption 3 *(a) Assumption 1 holds. (b) $\Psi(\theta, P)$ is four times continuously differentiable in $\mathcal{N}$. (c) $I - \Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P)$ is nonsingular. (d) $\Gamma(\theta, P) > 0$ for any $(a, x) \in A \times X$ and $(\theta, P) \in \Theta \times B_P$. (e) The operator $\phi_0^\Gamma(P) - P$ has a nonsingular Jacobian matrix at P^0.*

Assumption 3(c) is required for $\Gamma(\theta, P)$ to be well-defined. It would be possible to drop Assumption 3(d) by considering a trimmed version of $\Gamma(\theta, P)$, but for brevity we do not pursue it. As shown in Proposition 1, Assumption 3(e) is implied by $\rho(M_{\Gamma_\theta}\Gamma_P) < 1$ which holds when a sufficiently small value of δ is chosen.

Proposition 6 *Suppose Assumption 3 holds. Then (a) $\hat{P}_{RPM} - P^0 = O_p(n^{-1/2})$ and $n^{-1/2}(\hat{\theta}_{RPM} - \theta^0) \to_d N(0, V_{RPM})$, where $V_{RPM} = [\Omega_{\theta\theta}^\Gamma + \Omega_{\theta P}^\Gamma(I - \Gamma_P)^{-1}\Gamma_\theta]^{-1}\Omega_{\theta\theta}^\Gamma\{[\Omega_{\theta\theta}^\Gamma + \Omega_{\theta P}^\Gamma(I - \Gamma_P)^{-1}\Gamma_\theta]^{-1}\}'$. (b) Suppose we obtain $(\tilde{\theta}_j, \tilde{P}_j)$ from $\tilde{P}_{j-1}$ by the RPM algorithm. Then, there exists $c > 0$ such that $\tilde{\theta}_j - \hat{\theta}_{RPM} = O_p(||\tilde{P}_{j-1} - \hat{P}_{RPM}||)$ and $\tilde{P}_j - \hat{P}_{RPM} = M_{\Gamma_\theta}\Gamma_P(\tilde{P}_{j-1} - \hat{P}_{RPM}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{RPM}|| + ||\tilde{P}_{j-1} - \hat{P}_{RPM}||^2)$ uniformly in $\tilde{P}_{j-1} \in \mathcal{N}_P(c)$, where $M_{\Gamma_\theta} \equiv I - \Gamma_\theta(\Gamma_\theta'\Delta_P\Gamma_\theta)^{-1}\Gamma_\theta'\Delta_P$.*

Implementing the RPM algorithm is costly because it requires evaluating $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ for all the trial values of θ. We reduce the computational burden by evaluating $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ outside the optimization routine by using a preliminary estimate of θ. This modification has only a second-order effect on the convergence of the algorithm because the derivatives of $\Gamma(\theta, P)$ with respect to $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ are zero when evaluated at $P = \Psi(\theta, P)$; see the second term in (8). Let η be a preliminary estimate of θ. Replacing θ in $\Pi(\theta, P)$ and $\nabla_{P'}\Psi(\theta, P)$ with η, we define the following mapping:

$$\Gamma(\theta, P, \eta) \equiv \Psi(\theta, P) + [(I - \Pi(\eta, P)\nabla_{P'}\Psi(\eta, P)\Pi(\eta, P))^{-1} - I]\Pi(\eta, P)(\Psi(\theta, P) - P).$$

Once $\Pi(\eta, P)$ and $\nabla_{P'}\Psi(\eta, P)$ are computed, the computational cost of evaluating $\Gamma(\theta, P, \eta)$ across different values of θ would be similar to that of evaluating $\Psi(\theta, P)$.

Let $(\tilde{\theta}_0, \tilde{P}_0)$ be an initial estimator of (θ^0, P^0). For instance, $\tilde{\theta}_0$ can be the PML estimator. The *approximate RPM algorithm* iterates the following steps until $j = k$:

Step 1: Given $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\bar{\Theta}_j} n^{-1}\sum_{i=1}^n \ln \Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_i|x_i)$, where $\bar{\Theta}_j \equiv \{\theta \in \Theta : \Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a|x) \in [\xi, 1 - \xi] \text{ for all } (a, x) \in A \times X\}$ for an

arbitrary small $\xi > 0$. We impose this restriction in order to avoid computing $\ln(0)$.[16]

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

The following proposition shows that the approximate RPM algorithm achieves the same convergence rate as the original RPM algorithm in the first order. For $\epsilon > 0$, define a neighborhood of (θ^0, P^0) by $\mathcal{N}(\epsilon) = \{P : \max\{||\theta - \theta^0||, ||P - P^0||\} < \epsilon\}$.

Proposition 7 *Suppose Assumption 3 holds and we obtain* $(\tilde{\theta}_j, \tilde{P}_j)$ *from* $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ *by the approximate RPM algorithm. Then, there exists* $c > 0$ *such that* $\tilde{\theta}_j - \hat{\theta}_{RPM} = O_p(||\tilde{P}_{j-1} - \hat{P}_{RPM}|| + n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}||^2)$ *and* $\tilde{P}_j - \hat{P}_{RPM} = M_{\Gamma_\theta}\Gamma_P(\tilde{P}_{j-1} - \hat{P}_{RPM}) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{RPM}||^2 + n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{RPM}|| + ||\tilde{P}_{j-1} - \hat{P}_{RPM}||^2)$ *uniformly in* $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}(c)$.

By choosing δ sufficiently small, the dominant eigenvalue of $M_{\Gamma_\theta}\Gamma_P$ lies inside the unit circle, and the approximate RPM algorithm can converge to a consistent estimator even when the NPL algorithm diverges away from the true value. The following proposition states the local convergence of the approximate RPM algorithm when $\rho(M_{\Gamma_\theta}\Gamma_P) < 1$.

Proposition 8 *Suppose Assumption 3 holds,* $\rho(M_{\Gamma_\theta}\Gamma_P) < 1$, *and* $\{\tilde{\theta}_k, \tilde{P}_k\}$ *is generated by the approximate RPM algorithm starting from* $(\tilde{\theta}_0, \tilde{P}_0)$. *Then, there exists* $c_2 > 0$ *such that* $\Pr(\lim_{k\to\infty}(\tilde{\theta}_k, \tilde{P}_k) = (\hat{\theta}_{RPM}, \hat{P}_{RPM})) \to 1$ *as* $n \to \infty$ *if* $(\tilde{\theta}_0, \tilde{P}_0) \in \mathcal{N}(c_2)$.

We emphasize that implementing the approximate RPM algorithm is substantially more costly than the original NPL algorithm when the state space is large. This is because it requires computing the Jacobian matrix $\nabla_{P'}\Psi(\theta, P)$ and its eigenvalues at least once. In the supplementary appendix, we discuss how to implement the approximate RPM algorithm in detail, including how to further reduce the computational burden.[17]

4.3 The q-NPL algorithm

When the spectral radius of Λ_P or Ψ_P is smaller than but close to 1, the convergence of the NPL algorithm could be slow and the generated sequence could behave erratically. Furthermore, in such a case, the efficiency loss from using the NPL estimator compared to the MLE is substantial. To overcome these problems, consider a q-fold operator of Λ as

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

[16] In practice, we may consider a penalized objective function by truncating $\Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ so that it takes a value between ξ and $1 - \xi$, and adding a penalty term that penalizes θ such that $\Gamma(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \notin [\xi, 1 - \xi]$.

[17] In particular, one does not need to compute $\nabla_{P'}\Psi(\theta, P)$ and all its eigenvalues for every $(\tilde{\theta}_j, \tilde{P}_j)$. Given $(\tilde{\theta}_j, \tilde{P}_j, \tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, one can approximate $\Pi(\tilde{\theta}_j, \tilde{P}_j)\nabla_{P'}\Psi(\tilde{\theta}_j, \tilde{P}_j)\Pi(\tilde{\theta}_j, \tilde{P}_j)$ using $\Pi(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ and m finite differences of $\Psi(\tilde{\theta}_j, \tilde{P}_j)$.

We may define $\Gamma^q(\theta, P)$ and $\Psi^q(\theta, P)$ analogously. Define the *q-NPL (q-RPM) algorithm* by using a q-fold operator Λ^q, Γ^q, and Ψ^q in place of Λ, Γ, or Ψ in the original NPL (RPM) algorithm. In the following, we focus on Λ^q but the same argument applies to Γ^q and Ψ^q.

If the sequence of estimators generated by the q-NPL algorithm converges, its limit satisfies $\check{\theta} = \arg\max_{\theta\in\Theta} n^{-1}\sum_{i=1}^n \ln \Lambda^q(\theta, \check{P})(a_i|x_i)$ and $\check{\theta} = \Lambda^q(\check{\theta}, \check{P})$. Among the pairs $(\hat{\theta}, \hat{P})$ that satisfy these two conditions, the one that maximizes the value of the pseudo likelihood is called the *q-NPL estimator* and denoted by $(\hat{\theta}_{qNPL}, \hat{P}_{qNPL})$.

Since the result of Lemma 1 also applies here by replacing Ψ with Λ^q, the local convergence property of the q-NPL algorithm is primarily determined by the spectral radius of $\Lambda^q_P \equiv \nabla_{P'}\Lambda^q(\theta^0, P^0)$. When $\rho(\Lambda_P)$ is less than 1, the q-NPL algorithm converges faster than the NPL algorithm because $\rho(\Lambda^q_P) = (\rho(\Lambda_P))^q$. Moreover, the variance of the q-NPL estimator approaches that of the MLE as $q \to \infty$.

Applying the q-NPL algorithm, as defined above, is computationally intensive because the q-NPL Step 1 requires evaluating Λ^q at many different values of θ. We reduce the computational burden by introducing a linear approximation of $\Lambda^q(\theta, P)$ around (η, P), where η is a preliminary estimate of θ: $\Lambda^q(\theta, P, \eta) \equiv \Lambda^q(\eta, P) + \nabla_{\theta'}\Lambda^q(\eta, P)(\theta - \eta)$.

Given an initial estimator $(\tilde{\theta}_0, \tilde{P}_0)$, the *approximate q-NPL algorithm* iterates the following steps until $j = k$:

Step 1: Given $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$, update θ by $\tilde{\theta}_j = \arg\max_{\theta\in\Theta^q_j} n^{-1}\sum_{i=1}^n \ln \Lambda^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a_i|x_i)$, where $\Theta^q_j \equiv \{\theta \in \Theta : \tilde{\Lambda}^q(\theta, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(a|x) \in [\xi, 1-\xi] \text{ for all } (a, x) \in A \times X\}$ for an arbitrary small $\xi > 0$.

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

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 involves much fewer evaluations of $\Lambda(\theta, P)$ across different values of P and θ, compared to the original q-NPL algorithm.[18]

Define the q-NPL counterparts of $\tilde{\theta}_0(P)$, $\phi_0(P)$, and $\Omega_{\theta\theta}$ as $\tilde{\theta}^q_0(P) \equiv \arg\max_{\theta\in\Theta} E \ln \Lambda^q(\theta, P)(a_i|x_i)$, $\phi^q_0(P) = \Lambda^q(\tilde{\theta}^q_0(P), P)$, and $\Omega^q_{\theta\theta} \equiv E\nabla_\theta \ln \Lambda^q(\theta^0, P^0)(a_i|x_i)\nabla_{\theta'} \ln \Lambda^q(\theta^0, P^0)(a_i|x_i)$, respectively.

Assumption 4 *(a) Assumption 1 holds. (b) $\Psi(\theta, P)$ is four times continuously differentiable in $\mathcal{N}$. (c) There is a unique θ^0 such that $\Lambda^q(\theta^0, P^0) = P^0$. (d) $I - (\alpha\Psi_P + (1-\alpha)I)^q$ and $I - \Psi_P$ are nonsingular. (e) The operator $\phi^q_0(P) - P$ has a nonsingular Jacobian matrix at P^0.*

Assumption 4(c) is necessary for identifying θ^0 when the conditional probability is given by $\Lambda^q(\theta, P)$. This assumption rules out $\theta^1 \neq \theta^0$ that satisfies $\Lambda^q(\theta^1, P^0) = P^0$ even if $\Lambda(\theta^1, P^0) \neq P^0$. This occurs, for example, if $\Lambda(\theta^1, P^0) = P^1$ and $\Lambda(\theta^1, P^1) = P^0$ hold for $\theta^1 \neq \theta^0$ and

[18] Using one-sided numerical derivatives, evaluating $\nabla_{\theta'}\Lambda^q(\tilde{\theta}_j, \tilde{P}_j)$ requires $(K+1)q$ function evaluations of $\Psi(\theta, P)$.

$P^1 \neq P^0$. Assumption 4(d) is necessary for $\Omega^q_{\theta\theta}$ to be nonsingular. Since $\Lambda^q_P = (\alpha\Psi_P + (1-\alpha)I)^q$, the first condition holds if $\rho(\Lambda^q_P) < 1$ from 19.15 of Seber (2007).

The following proposition establishes that asymptotics of the q-NPL estimator and the convergence property of the approximate q-NPL algorithm. Proposition 9(c) implies that, when q is sufficiently large, the q-NPL estimator is more efficient than the NPL estimator, provided that additional conditions in Assumption 4 hold. Proposition 9(d) corresponds to Lemma 2.

Proposition 9 *Suppose that Assumption 4 holds. Then (a)* $\hat{P}_{qNPL} - P^0 = O_p(n^{-1/2})$ *and* $n^{-1/2}(\hat{\theta}_{qNPL} - \theta^0) \to_d N(0, V_{qNPL})$, *where* $V_{qNPL} = [\Omega^q_{\theta\theta} + \Omega^q_{\theta P}(I - \Lambda_P)^{-1}\Lambda^q_\theta]^{-1}\Omega^q_{\theta\theta}\{[\Omega^q_{\theta\theta} + \Omega^q_{\theta P}(I - \Lambda_P)^{-1}\Lambda^q_\theta]^{-1}\}'$. *(b) Suppose we obtain* $(\tilde{\theta}_j, \tilde{P}_j)$ *from* $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1})$ *by the approximate q-NPL algorithm. Then, there exists* $c > 0$ *such that* $\tilde{\theta}_j - \hat{\theta}_{qNPL} = O_p(||\tilde{P}_{j-1} - \hat{P}_{qNPL}|| + n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2)$ *and* $\tilde{P}_j - \hat{P}_{qNPL} = M_{\Lambda^q_\theta}\Lambda^q_P(\tilde{P}_{j-1} - \hat{P}_{qNPL}) + O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}_{qNPL}||^2 + n^{-1/2}||\tilde{P}_{j-1} - \hat{P}_{qNPL}|| + ||\tilde{P}_{j-1} - \hat{P}_{qNPL}||^2)$ *uniformly in* $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}(c)$, *where* $M_{\Lambda^q_\theta} \equiv I - \Lambda^q_\theta((\Lambda^q_\theta)'\Delta_P\Lambda^q_\theta)^{-1}(\Lambda^q_\theta)'\Delta_P$ *with* $\Lambda^q_\theta \equiv \nabla_{\theta'}\Lambda^q(\theta^0, P^0)$. *(c) If* $\rho(\Lambda_P) < 1$, *then* $V_{qNPL} \to V_{MLE}$ *as* $q \to \infty$. *(d) Suppose* $\{\tilde{\theta}_k, \tilde{P}_k\}$ *is generated by the approximate q-NPL algorithm starting from* $(\tilde{\theta}_0, \tilde{P}_0)$ *and* $\rho(M_{\Lambda^q_\theta}\Lambda^q_P) < 1$. *Then, there exists* $c_2 > 0$ *such that* $\Pr(\lim_{k\to\infty}(\tilde{\theta}_k, \tilde{P}_k) = (\hat{\theta}_{qNPL}, \hat{P}_{qNPL})) \to 1$ *as* $n \to \infty$ *if* $(\tilde{\theta}_0, \tilde{P}_0) \in \mathcal{N}(c_2)$.

5 Monte Carlo experiments

We consider a dynamic game model of market entry and exit studied in Section 4 of AM07. We set the number of firms $N = 3$. The profit of firm i operating in market m in period t is equal to $\tilde{\Pi}_{it}(1) = \theta_{RS}\ln S_{mt} - \theta_{RN}\ln(1 + \sum_{j\neq i} a_{jmt}) - \theta_{FC,i} - \theta_{EC}(1 - a_{im,t-1}) + \epsilon_{imt}(1)$, whereas its profit is $\tilde{\Pi}_{it}(0) = \epsilon_{imt}(0)$ if the firm is not operating. We assume that $\{\epsilon_{imt}(0), \epsilon_{imt}(1)\}$ follow i.i.d. type I extreme value distribution, and S_{mt} follows an exogenous first-order Markov process $f_S(S_{m,t+1}|S_{mt})$.[19] The discount factor is set to $\beta = 0.96$, and the parameter values are given by $\theta_{RS} = 1.0$, $\theta_{EC} = 1.0$, $\theta_{FC,1} = 1.0$, $\theta_{FC,2} = 0.9$, and $\theta_{FC,3} = 0.8$. The parameter θ_{RN} determines the degree of strategic substitutabilities among firms and is the main determinant of the dominant eigenvalue of Ψ_P. All of the eigenvalues of Ψ_P are inside the unit circle for $\theta_{RN} = 1$ and 2 while the smallest eigenvalues are less than -1 for $\theta_{RN} = 4$ and 6. We therefore let θ_{RN} take on a value of 2 or 4 across experiments and examine the performance of different estimators. We estimate θ_{RS} and θ_{RN}, leaving the other parameters fixed at the true values.

To generate an observation, we first randomly draw $x_m = \{S_{m1}, a_{1m0}, a_{2m0}, a_{3m0}\}$ from the steady-state distribution implied by the model. Then, given x_m, we draw $\{a_{1m1}, a_{2m1}, a_{3m1}\}$

[19] The state space for the market size S_{mt} is $\{2, 6, 10\}$. The transition probability matrix of S_{mt} is given by

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

from the equilibrium conditional choice probabilities. We replicate 1000 simulated samples for each of $n = 500$, 2000, and 8000 observations.

As shown in Table 1, the spectral radius of $M_{\Psi_\theta}\Psi_P$ and $M_{\Lambda_\theta}\Lambda_P$ is very similar to that of Ψ_P and Λ_P, respectively. Thus, in view of Lemma 1, the convergence property of the NPL algorithm is primarily determined by the dominant eigenvalue of Ψ_P and Λ_P.

The first panel of Table 2 compares the performance of sequential estimators generated by the following four sequential algorithms evaluated at $k = 50$ iterations: (i) the NPL algorithm using Ψ, (ii) the NPL algorithm using Λ, (iii) the approximate RPM algorithm using $\Gamma(\theta, P, \eta)$ with $\delta = 0.5$, and (iv) the approximate q-NPL using $\Lambda^q(\theta, P, \eta)$ with $q = 4$. They are denoted by "NPL-Ψ," "NPL-Λ," "RPM," and "q-NPL-Λ^q," respectively. The second panel of Table 2 reports the performance of two-step (PML) version of the above four estimators. These estimators are included for reference; they do not need iteration but require a root-n consistent initial nonparametric estimate of P. They are denoted by "PML-Ψ," "PML-RPM," and "PML-Λ^q," respectively.[20] We report the bias and the root mean squared errors (RMSE, henceforth) of $\hat{\theta}_{RN}$ and $\hat{\theta}_{RS}$ across different estimators.

For $\theta_{RN} = 2$, the NPL-Ψ has substantially improved performance over the PML-Ψ across different sample sizes, and the NPL-Λ and NPL-Ψ converge to the same estimate. On the other hand, when $\theta_{RN} = 4$ the NPL-Ψ performs substantially worse than the NPL-Λ, reflecting divergence. Further, as the sample size increases from 500 to 8000, the RMSE of the NPL-Λ decreases approximately at the rate of $n^{1/2}$, but the RMSE of the NPL-Ψ decreases at a much slower rate. For $\theta_{RN} = 4$ and $n = 8000$, the RMSE of the NPL-Ψ is even larger than that of the PML-Ψ.

Across different sample sizes and parameters, the RPM and the q-NPL-Λ^q outperform the NPL-Ψ. The PML-RPM and the PML-Λ^q also perform substantially better than the PML-Ψ, suggesting that our proposed alternative sequential methods are useful even when the researcher wants to make just one NPL iteration rather than iterate the NPL algorithm until convergence.

The first four rows of Table 3 compare the RMSE across the estimators of θ_{RN} generated by different algorithms after $k = 2, 5, 10, \ldots, 25$ iterations when $n = 8000$. For $\theta_{RN} = 2$, the RMSE changes little after $j = 5$ iterations across all the algorithms, indicating their convergence. For $\theta_{RN} = 4$, the RMSE of the NPL-Ψ sequence increases with the number of iterations whereas our proposed estimators converge after 10 iterations. The last two rows of Table 3 report the RMSE of the first and the second differences of the NPL-Ψ sequence in order to examine its possible convergence to a 2-period cycle. When $\theta_{RN} = 4$, the NPL-Ψ sequence does not converge to a NPL fixed point but they gradually converge *every other iteration*, suggesting its convergence toward a 2-period cycle.

[20] We do not report PML-Λ because it is identical to PML-Ψ. See the paragraph following Proposition 4. The PML-RPM and the PML-Λ^q take one RPM or approximate q-NPL step from the original PML estimator with Ψ and, thus, they are three step estimators. Their asymptotic properties can be easily derived from Proposition 1 of AM07, apart from changes in regularity conditions.

6 Concluding remarks and extension

This paper analyzes the convergence properties of the NPL algorithm to estimate a class of structural models characterized by a fixed point constraint. We show that, when the fixed point mapping has a local contraction property, the NPL algorithm converges to a consistent estimator if started from a neighborhood of the true value.

In practice, the convergence condition may be violated. In such a case, the NPL algorithm will not converge to a consistent estimator even if it is started from a neighborhood of the true parameter value. We develop alternative sequential estimators that can be used even when the original fixed point mapping is not locally contractive. As our simulations illustrate, these alternative estimators work well even when the original fixed point mapping is not a contraction, and their performance is substantially better than that of the two-step PML estimator.

Our convergence analysis is local. In a model with multiple NPL fixed points, whether the sequential algorithms analyzed in this paper can be used to obtain a consistent NPL fixed point depends on the initial value of P. Thus, when a reliable initial estimate is not available, it is recommended to repeatedly apply the NPL algorithm with different initial values. A closely related unresolved issue is the size of the domain of attraction for these sequential algorithms. For instance, if the q-NPL algorithm has a smaller domain of attraction than the NPL algorithm, then the finite sample properties of the q-NPL estimator may be worse than those of the NPL estimator. Examining such a possibility is an important future topic.

In the supplementary appendix, we discuss further results including models with permanent unobserved heterogeneity, sequential generalized method of moments estimators, an approximate fixed point algorithm, and additional Monte Carlo results.

References

Aiyagari, S. Rao (1994). "Uninsured idiosyncratic risk and aggregate saving." *Quarterly Journal of Economics* 109(3): 659-684.

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, H. Hong, and D. Nekipelov (2009). Nonparametric and semiparametric analysis of a dynamic discrete game. Mimeographed, Stanford university.

Başar, T. (1987). "Relaxation Techniques and Asynchronous Algorithms for On-line Computation of Noncooperative Equilibria." *Journal of Economic Dynamics and Controls*, 11: 531-549.

Chu, K-W E. (1990). "On multiple eigenvalues of matrices depending on several parameters," *SIAM Journal of Numerical Analysis* 27(5): 1368-1385.

Horn R. A. and C. R. Johnson (1985) *Matrix Analysis*. Cambridge University Press.

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

Judd, L. J. (1998) *Numerical Methods in Economics*. Cambridge, Massachusetts: The MIT Press.

Kasahara, H. and K. Shimotsu (2008) "Pseudo-likelihood Estimation and Bootstrap Inference for Structural Discrete Markov Decision Models," *Journal of Econometrics*, 146: 92-106.

Kasahara, H. and K. Shimotsu (2009) "Nonparametric identification of finite mixture models of dynamic discrete choices," *Econometrica*, 77(1): 135-175.

Krusell, P. and A. Smith Jr. (1998) "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, 106(5): 867-896

Ljungqvist, L. and T. J. Sargent (2004) *Recursive Macroeconomic Theory*, 2nd ed., MIT Press.

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 (2008). "Asymptotic least squares estimators for dynamic games," *Review of Economic Studies*, 75, 901-928.

Pesendorfer, M. and P. Schmidt-Dengler (2009). "Sequential estimation of dynamic discrete games: a comment," *Econometrica*, forthcoming.

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

Seber, G. A. F. (2007). *A Matrix Handbook for Statisticians*. Wiley.

Shroff, G. M. and H. B. Keller (1993) "Stabilization of unstable procedures: the recursive projection method," *SIAM Journal of Numerical Analysis*, 30(4): 1099-1120.

Strogatz, S. H. (1994). *Nonlinear Dynamics And Chaos: With Applications To Physics, Biology, Chemistry And Engineering*. Westview Press

Su, C.-L. and K. L. Judd (2008) Constrained optimization approaches to estimation of structural models. Mimeographed, University of Chicago.

Table 1: The Spectral Radius of Ψ_P and Λ_P

θ_{RN}	α	$\rho(\Psi_P)$	$\rho(\Lambda_P)$	$\rho(M_{\Psi_\theta}\Psi_P)$	$\rho(M_{\Lambda_\theta}\Lambda_P)$
1	0.9407	0.3365	0.2572	0.2916	0.2557
2	0.8830	0.6925	0.4945	0.5949	0.4936
4	0.8250	1.1839	0.8017	1.1799	0.8046
6	0.7730	1.4788	0.9161	1.4777	0.9153

The second column reports the optimal choice of α under which Λ_P has the smallest spectral radius.

Table 2: Bias and RMSE

	Estimator	$\theta_{RN}=2$						$\theta_{RN}=4$					
		$n=500$		$n=2000$		$n=8000$		$n=500$		$n=2000$		$n=8000$	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\hat{\theta}_{RS}$	NPL-Ψ	-0.0151	0.1347	-0.0002	0.0660	-0.0023	0.0323	-0.0095	0.0676	-0.0062	0.0490	-0.0005	0.0408
	NPL-Λ	-0.0151	0.1347	-0.0002	0.0660	-0.0023	0.0323	0.0028	0.0575	-0.0006	0.0294	-0.0003	0.0143
	RPM	-0.0174	0.1331	-0.0028	0.0642	-0.0027	0.0320	0.0029	0.0576	-0.0012	0.0284	0.0000	0.0136
	q-NPL-Λ^q	-0.0117	0.1240	0.0002	0.0606	-0.0018	0.0305	0.0015	0.0542	-0.0009	0.0277	0.0000	0.0136
	PML-Ψ	-0.2215	0.2698	-0.0717	0.1112	-0.0229	0.0474	-0.1280	0.1557	-0.0341	0.0514	-0.0082	0.0207
	PML-RPM	0.1353	0.2380	0.0658	0.1072	0.0203	0.0403	0.1166	0.1823	0.0211	0.0457	0.0043	0.0176
	PML-Λ^q	-0.0133	0.1475	0.0016	0.0629	-0.0018	0.0307	0.0142	0.0783	-0.0035	0.0290	-0.0003	0.0141
$\hat{\theta}_{RN}$	NPL-Ψ	-0.0467	0.4705	-0.0009	0.2339	-0.0095	0.1130	-0.1417	0.2572	-0.1414	0.2314	-0.0918	0.1612
	NPL-Λ	-0.0467	0.4705	-0.0009	0.2339	-0.0095	0.1130	0.0241	0.1424	-0.0001	0.0739	0.0013	0.0352
	RPM	-0.0544	0.4642	-0.0102	0.2274	-0.0111	0.1116	0.0249	0.1604	-0.0003	0.0841	0.0014	0.0342
	q-NPL-Λ^q	-0.0358	0.4280	0.0002	0.2131	-0.0079	0.1052	0.0228	0.1351	0.0000	0.0690	0.0014	0.0328
	PML-Ψ	-0.7895	0.9604	-0.2565	0.3949	-0.0828	0.1687	-0.7713	0.9094	-0.1964	0.2599	-0.0462	0.0937
	PML-RPM	0.4523	0.8255	0.2232	0.3754	0.0687	0.1401	0.6101	0.7821	0.1282	0.1848	0.0335	0.0600
	PML-Λ^q	-0.0603	0.5177	0.0021	0.2215	-0.0083	0.1061	0.1619	0.2704	0.0044	0.0745	0.0035	0.0366

Table 3: RMSE of $\hat{\theta}_{RN,k}$ for $k = 2, 5, 10, ..., 25$ at $n = 8000$

		$\theta_{RN}=2$						$\theta_{RN}=4$					
		$k=2$	$k=5$	$k=10$	$k=15$	$k=20$	$k=25$	$k=2$	$k=5$	$k=10$	$k=15$	$k=20$	$k=25$
$\tilde{\theta}_{RN,k}$	NPL-Ψ	0.1196	0.1133	0.1130	0.1130	0.1130	0.1130	0.0713	0.0748	0.0807	0.1235	0.1299	0.1593
	NPL-Λ	0.1227	0.1131	0.1130	0.1130	0.1130	0.1130	0.0651	0.0363	0.0353	0.0352	0.0352	0.0352
	RPM	0.1401	0.1122	0.1120	0.1118	0.1117	0.1116	0.0600	0.0357	0.0350	0.0341	0.0343	0.0342
	q-NPL-Λ^q	0.1061	0.1051	0.1052	0.1052	0.1052	0.1052	0.0366	0.0332	0.0328	0.0328	0.0328	0.0328
RMSE of $(\tilde{\theta}_{RN,k+1} - \tilde{\theta}_{RN,k})$		0.0532	0.0041	0.0003	0.0000	0.0000	0.0000	0.1272	0.1106	0.1551	0.2037	0.2410	0.2624
RMSE of $(\tilde{\theta}_{RN,k+2} - \tilde{\theta}_{RN,k})$		0.0505	0.0017	0.0001	0.0000	0.0000	0.0000	0.0310	0.0152	0.0157	0.0132	0.0101	0.0076

The last two rows report the RMSE of $(\tilde{\theta}_{RN,k+1} - \tilde{\theta}_{RN,k})$ and $(\tilde{\theta}_{RN,k+2} - \tilde{\theta}_{RN,k})$ for NPL-Ψ.

7 Appendix: Proofs

Throughout the proofs, let "wpa1" abbreviate "with probability approaching one as $n \to \infty$." The $O_p()$ terms in the proof such as $O_p(||\tilde{P}_{j-1} - \hat{P}_{NPL}||)$ are uniform, but we suppress the reference to their uniformity for brevity.

7.1 Proof of Lemma 1

We suppress the subscript NPL from $\hat{P}_{NPL}$ and $\hat{\theta}_{NPL}$. Define $\overline{\psi}(\theta, P) \equiv n^{-1} \sum_{i=1}^{n} \ln \Psi(\theta, P)(a_i|x_i)$ and $\psi(\theta, P) \equiv E \ln \Psi(\theta, P)(a_i|x_i)$. For $\epsilon > 0$, define a neighborhood $\mathcal{N}(\epsilon) = \{(\theta, P) : \max\{||\theta - \theta^0||, ||P - P^0||\} < \epsilon\}$. Then, there exists $\epsilon_1 > 0$ such that $\mathcal{N}(\epsilon_1) \subset \mathcal{N}$ and $\sup_{(\theta,P)\in\mathcal{N}(\epsilon_1)} ||\nabla_{\theta\theta'}\psi(\theta, P)^{-1}|| < \infty$ because $\nabla_{\theta\theta'}\psi(\theta, P)$ is continuous and $\nabla_{\theta\theta'}\psi(\theta^0, P^0)$ is nonsingular.

First, we assume $(\tilde{\theta}_j, \tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ and derive the stated representation of $\tilde{\theta}_j - \hat{\theta}$ and $\tilde{P}_j - \hat{P}$. We later show $(\tilde{\theta}_j, \tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ wpa1 if c is taken sufficiently small. The first order condition for $\tilde{\theta}_j$ is $\nabla_\theta \overline{\psi}(\tilde{\theta}_j, \tilde{P}_{j-1}) = 0$. Expanding it around $(\hat{\theta}, \hat{P})$ and using $\nabla_\theta \overline{\psi}(\hat{\theta}, \hat{P}) = 0$ gives

$$0 = \nabla_{\theta\theta'}\overline{\psi}(\bar{\theta}, \bar{P})(\tilde{\theta}_j - \hat{\theta}) + \nabla_{\theta P'}\overline{\psi}(\bar{\theta}, \bar{P})(\tilde{P}_{j-1} - \hat{P}), \tag{9}$$

where $(\bar{\theta}, \bar{P})$ lie between $(\tilde{\theta}_j, \tilde{P}_{j-1})$ and $(\hat{\theta}, \hat{P})$. Write (9) as $\tilde{\theta}_j - \hat{\theta} = -\nabla_{\theta\theta'}\overline{\psi}(\bar{\theta}, \bar{P})^{-1}\nabla_{\theta P'}\overline{\psi}(\bar{\theta}, \bar{P})(\tilde{P}_{j-1} - \hat{P})$, then the stated uniform bound of $\tilde{\theta}_j - \hat{\theta}$ follows because (i) $(\bar{\theta}, \bar{P}) \in \mathcal{N}(\epsilon_1)$ wpa1 since $(\tilde{\theta}_j, \tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ and $(\hat{\theta}, \hat{P})$ is consistent, and (ii) $\sup_{(\theta,P)\in\mathcal{N}(\epsilon_1)} ||\nabla_{\theta\theta'}\overline{\psi}(\theta, P)^{-1}\nabla_{\theta P'}\overline{\psi}(\theta, P)|| = O_p(1)$ since $\sup_{(\theta,P)\in\mathcal{N}(\epsilon_1)} ||\nabla_{\theta\theta'}\psi(\theta, P)^{-1}|| < \infty$ and $\sup_{(\theta,P)\in\mathcal{N}} ||\nabla^2\overline{\psi}(\theta, P) - \nabla^2\psi(\theta, P)|| = o_p(1)$, where the latter follows from Lemma 2.4 of Newey and McFadden (1994).

For the bound of $\tilde{P}_j - \hat{P}$, first we collect the following results, which follow from the Taylor expansion around (θ^0, P^0), root-n consistency of $(\hat{\theta}, \hat{P})$, and the information matrix equality.

$$\begin{array}{ll} \nabla_{\theta\theta'}\overline{\psi}(\hat{\theta}, \hat{P}) = -\Omega_{\theta\theta} + O_p(n^{-1/2}), & \nabla_{\theta P'}\overline{\psi}(\hat{\theta}, \hat{P}) = -\Omega_{\theta P} + O_p(n^{-1/2}), \\ \nabla_{\theta'}\Psi(\hat{\theta}, \hat{P}) = \Psi_\theta + O_p(n^{-1/2}), & \nabla_{P'}\Psi(\hat{\theta}, \hat{P}) = \Psi_P + O_p(n^{-1/2}). \end{array} \tag{10}$$

Expand the right hand side of $\tilde{P}_j = \Psi(\tilde{\theta}_j, \tilde{P}_{j-1})$ twice around $(\hat{\theta}, \hat{P})$ and use $\Psi(\hat{\theta}, \hat{P}) = \hat{P}$ and $\tilde{\theta}_j - \hat{\theta} = O_p(||\tilde{P}_{j-1} - \hat{P}||)$, then we obtain $\tilde{P}_j - \hat{P} = \nabla_{\theta'}\Psi(\hat{\theta}, \hat{P})(\tilde{\theta}_j - \hat{\theta}) + \nabla_{P'}\Psi(\hat{\theta}, \hat{P})(\tilde{P}_{j-1} - \hat{P}) + O_p(||\tilde{P}_{j-1} - \hat{P}||^2)$ since $\sup_{(\theta,P)\in\mathcal{N}(\epsilon_1)} \nabla^3\Psi(\theta, P) < \infty$. Applying (10) and $\tilde{\theta}_j - \hat{\theta} = O_p(||\tilde{P}_{j-1} - \hat{P}||)$ to the right hand side gives

$$\tilde{P}_j - \hat{P} = \Psi_\theta(\tilde{\theta}_j - \hat{\theta}) + \Psi_P(\tilde{P}_{j-1} - \hat{P}) + O_p(||\tilde{P}_{j-1} - \hat{P}||^2) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}||). \tag{11}$$

We proceed to refine (9) to write $\tilde{\theta}_j - \hat{\theta}$ in terms of $\tilde{P}_{j-1} - \hat{P}$ and substitute it into (11). Expanding $\nabla_{\theta\theta'}\overline{\psi}(\bar{\theta}, \bar{P})$ in (9) around $(\hat{\theta}, \hat{P})$, noting that $||\bar{\theta} - \hat{\theta}|| \leq ||\tilde{\theta}_j - \hat{\theta}||$ and $||\bar{P} - \hat{P}|| \leq ||\tilde{P}_{j-1} - \hat{P}||$, and using $\tilde{\theta}_j - \hat{\theta} = O_p(||\tilde{P}_{j-1} - \hat{P}||)$, we obtain $\nabla_{\theta\theta'}\overline{\psi}(\bar{\theta}, \bar{P}) = \nabla_{\theta\theta'}\overline{\psi}(\hat{\theta}, \hat{P}) + O_p(||\tilde{P}_{j-1} - \hat{P}||)$. Further, applying (10) gives $\nabla_{\theta\theta'}\overline{\psi}(\bar{\theta}, \bar{P}) = -\Omega_{\theta\theta} + O_p(n^{-1/2}) + O_p(||\tilde{P}_{j-1} - \hat{P}||)$. Similarly, we

obtain $\nabla_{\theta P'}\overline{\psi}(\bar{\theta},\bar{P}) = -\Omega_{\theta P} + O_p(n^{-1/2}) + O_p(||\tilde{P}_{j-1} - \hat{P}||)$. Using these results, refine (9) as $\tilde{\theta}_j - \hat{\theta} = -\Omega_{\theta\theta}^{-1}\Omega_{\theta P}(\tilde{P}_{j-1} - \hat{P}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}|| + ||\tilde{P}_{j-1} - \hat{P}||^2)$. Substituting this into (11) in conjunction with $\Omega_{\theta\theta}^{-1}\Omega_{\theta P} = (\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P\Psi_P$ gives the stated result.

It remains to show $(\tilde{\theta}_j, \tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$ wpa1 if c is taken sufficiently small. Let $\mathcal{N}_\theta(\epsilon_1) \equiv \{\theta : ||\theta - \theta^0|| < \epsilon_1\}$ and define $\Delta = \psi(\theta^0, P^0) - \sup_{\theta \in \mathcal{N}_\theta(\epsilon_1)^c \cap \Theta} \psi(\theta, P^0) > 0$, where the last inequality follows from information inequality, compactness of $\mathcal{N}_\theta(\epsilon_1)^c \cap \Theta$, and continuity of $\psi(\theta, P)$. It follows that $\Pr(\tilde{\theta}_j \notin \mathcal{N}_\theta(\epsilon_1)) \leq \Pr(\psi(\theta^0, P^0) - \psi(\tilde{\theta}_j, P^0) \geq \Delta)$. Further, observe that $\psi(\theta^0, P^0) - \psi(\tilde{\theta}_j, P^0) \leq \overline{\psi}(\theta^0, \tilde{P}_{j-1}) - \overline{\psi}(\tilde{\theta}_j, \tilde{P}_{j-1}) + 2\sup_{\theta\in\Theta} |\psi(\theta, P^0) - \psi(\theta, \tilde{P}_{j-1})| + 2\sup_{(\theta,P)\in\Theta\times B_P} |\overline{\psi}(\theta, P) - \psi(\theta, P)| \leq 2\sup_{\theta\in\Theta} |\psi(\theta, P^0) - \psi(\theta, \tilde{P}_{j-1})| + 2\sup_{(\theta,P)\in\Theta\times B_P} |\overline{\psi}(\theta, P) - \psi(\theta, P)|$, where the second inequality follows from the definition of $\tilde{\theta}_j$. From continuity of $\psi(\theta, P)$, there exists $\epsilon_2(\Delta) > 0$ such that the first term on the right is smaller than $\Delta/2$ if $\epsilon \leq \epsilon_2(\Delta)$. The second term on the right is $o_p(1)$ from Lemma 2.4 of Newey and McFadden (1994). Hence, $\Pr(\tilde{\theta}_j \notin \mathcal{N}_\theta(\epsilon_1)) \to 0$ if $\epsilon \leq \epsilon_2(\Delta)$, and setting $c \leq \min\{\epsilon_1, \epsilon_2(\Delta)\}$ gives $\Pr((\tilde{\theta}_j, \tilde{P}_{j-1}) \notin \mathcal{N}(\epsilon_1)) \to 0$. □

7.2 Proof of Lemma 2

We suppress the subscript NPL from $\hat{P}_{NPL}$. Let $b > 0$ be a constant such that $\rho(M_{\Psi_\theta}\Psi_P) + 2b < 1$. From Lemma 5.6.10 of Horn and Johnson (1985), there is a matrix norm $||\cdot||_\alpha$ such that $||M_{\Psi_\theta}\Psi_P||_\alpha \leq \rho(M_{\Psi_\theta}\Psi_P) + b$. Define a vector norm $||\cdot||_\beta$ for $x \in \mathbb{R}^L$ as $||x||_\beta = |||[x\ 0 \dots 0]||_\alpha$, then a direct calculation gives $||Ax||_\beta = ||A[x\ 0 \dots 0]||_\alpha \leq ||A||_\alpha||x||_\beta$ for any matrix A. From the equivalence of vector norms in $\mathbb{R}^L$ (see, for example, Corollary 5.4.5 of Horn and Johnson (1985)), we can restate Lemma 1 in terms of $||\cdot||_\beta$ as follows: there exists $c > 0$ such that $\tilde{P}_j - \hat{P} = M_{\Psi_\theta}\Psi_P(\tilde{P}_{j-1} - \hat{P}) + O_p(n^{-1/2}||\tilde{P}_{j-1} - \hat{P}||_\beta + ||\tilde{P}_{j-1} - \hat{P}||_\beta^2)$ holds uniformly in $\tilde{P}_{j-1} \in \{P : ||P - P^0||_\beta < c\}$. We rewrite this statement further so that it is amenable to recursive substitution. First, note that $||M_{\Psi_\theta}\Psi_P(\tilde{P}_{j-1} - \hat{P})||_\beta \leq ||M_{\Psi_\theta}\Psi_P||_\alpha||\tilde{P}_{j-1} - \hat{P}||_\beta \leq (\rho(M_{\Psi_\theta}\Psi_P) + b)||\tilde{P}_{j-1} - \hat{P}||_\beta$. Second, rewrite the O_p term as $O_p(n^{-1/2} + ||\tilde{P}_{j-1} - \hat{P}||_\beta)||\tilde{P}_{j-1} - \hat{P}||_\beta$. Set $c < b$, then this term is smaller than $b||\tilde{P}_{j-1} - \hat{P}||_\beta$ wpa1. Third, since $\hat{P}$ is consistent, $\{P : ||P - \hat{P}||_\beta < c/2\} \subset \{P : ||P - P^0||_\beta < c\}$ wpa1. Consequently, $||\tilde{P}_j - \hat{P}||_\beta \leq (\rho(M_{\Psi_\theta}\Psi_P) + 2b)||\tilde{P}_{j-1} - \hat{P}||_\beta$ holds wpa1 for all $\tilde{P}_{j-1}$ in $\{P : ||P - \hat{P}||_\beta < c/2\}$. Because each NPL updating of (θ, P) uses the same pseudo-likelihood function, we may recursively substitute for the $\tilde{P}_j$'s, and hence $\lim_{k\to\infty} \tilde{P}_k = \hat{P}$ wpa1 if $||\tilde{P}_0 - \hat{P}||_\beta < c/2$. The stated result follows from applying the equivalence of vector norms in $\mathbb{R}^L$ to $||\tilde{P}_0 - \hat{P}||_\beta$ and $||\tilde{P}_0 - \hat{P}||$ and using the consistency of $\hat{P}$. □

7.3 Proof of Proposition 1

Differentiating $\phi_0(P)$ gives

$$\nabla_{P'}\phi_0(P^0) = \nabla_{\theta'}\Psi(\tilde{\theta}_0(P^0), P^0)\nabla_{P'}\tilde{\theta}_0(P^0) + \nabla_{P'}\Psi(\tilde{\theta}_0(P), P^0). \tag{12}$$

We proceed to derive a representation of $\nabla_{P'}\tilde{\theta}_0(P^0)$ and substitute it into (12). The first order condition for $\tilde{\theta}_0(P)$ implies $\nabla_\theta E \ln \Psi(\tilde{\theta}_0(P), P)(a_i|x_i) = 0$. Taking its derivative with respect to P gives $\nabla_{\theta\theta'} E \ln \Psi(\tilde{\theta}_0(P), P)(a_i|x_i)\nabla_{P'}\tilde{\theta}_0(P) + \nabla_{\theta P'} E \ln \Psi(\tilde{\theta}_0(P), P)(a_i|x_i) = 0$. Evaluating this at P^0 and using $\tilde{\theta}(P^0) = \theta^0$, we obtain $\nabla_{P'}\tilde{\theta}(P^0) = -(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P\Psi_P$. Substituting this representation of $\nabla_{P'}\tilde{\theta}_0(P^0)$ into (12) and using $\tilde{\theta}(P^0) = \theta^0$, we obtain $\nabla_{P'}\phi_0(P^0) = (I - \Psi_\theta(\Psi_\theta'\Delta_P\Psi_\theta)^{-1}\Psi_\theta'\Delta_P)\Psi_P = M_{\Psi_\theta}\Psi_P$. Therefore, the Jacobian of $\phi_0(P) - P$ at P^0 equals $M_{\Psi_\theta}\Psi_P - I$. From 19.15 of Seber (2007), $M_{\Psi_\theta}\Psi_P - I$ is nonsingular if $\rho(M_{\Psi_\theta}\Psi_P) < 1$. □

7.4 Proof of Proposition 2

First, note that $\tilde{P}_j$ for $j \geq 1$ satisfies restriction (3) because it is generated by $\Psi(\theta, P)$. The restrictions (3)–(4) do not affect the validity of Lemma 1 because (i) the fixed point constraint in terms of $\Psi(\theta, P)$ and of $\Psi^+(\theta, P^+)$ are equivalent, and (ii) the restrictions (3)–(4) do not affect the order of magnitude of the derivatives of $\Psi(\theta, P)$.

For the updating formula of P^+ and P^-, taking the derivative of (4) gives

$$\nabla_{P'}\Psi(\theta, P) = \begin{pmatrix} \nabla_{P^{+\prime}}\Psi^+(\theta, P^+) & 0 \\ -\mathcal{E}\nabla_{P^{+\prime}}\Psi^+(\theta, P^+) & 0 \end{pmatrix} = \begin{pmatrix} U\nabla_{P^{+\prime}}\Psi^+(\theta, P^+) & 0 \end{pmatrix}. \tag{13}$$

Substituting this into $M_{\Psi_\theta}\Psi_P$, using $\Psi_\theta = U\Psi_\theta^+$, and rearranging terms give $M_{\Psi_\theta}\Psi_P = [UM_{\Psi_\theta}^+\Psi_{P^+}^+ \vdots 0]$, and the stated updating formula follows. The equivalence of the eigenvalues follows from $\det(M_{\Psi_\theta}\Psi_P - \lambda I_{\dim(P)}) = \det(M_{\Psi_\theta}^+\Psi_{P^+}^+ - \lambda I_{\dim(P^+)})\det(-\lambda I_{\dim(P^-)})$ and $\det(\Psi_P - \lambda I_{\dim(P)}) = \det(\Psi_{P^+}^+ - \lambda I_{\dim(P^+)})\det(-\lambda I_{\dim(P^-)})$. □

7.5 Proof of Proposition 3

Observe that, with the simplex restriction on P, $\Psi(\theta, P)$ takes the form

$$\Psi(\theta, P) = \begin{pmatrix} M_1(\theta) & M_2(\theta) \end{pmatrix} \begin{pmatrix} P^+ \\ 1 - \mathbf{1}_{L-1}'P^+ \end{pmatrix} = M_1(\theta)P^+ + M_2(\theta)(1 - \mathbf{1}_{L-1}'P^+).$$

Then part (a) follows straightforwardly.

For part (b), partition $M(\theta)$ and define a matrix E as

$$M(\theta) = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}, \quad E \equiv \begin{pmatrix} I_{L-1} & \mathbf{0} \\ -\mathbf{1}_{L-1}' & 1 \end{pmatrix},$$

where we suppress θ from the $M_{ij}(\theta)$'s, M_{11} is $(L-1) \times (L-1)$, and I_k is a k-dimensional

identity matrix. Direct calculation and noting that $\nabla_{P^{+\prime}}\Psi^{+}(\theta,P^{+}) = M_{11} - M_{12}\mathbf{1}'_{L-1}$ gives

$$(M(\theta)-\lambda I_L)E = \begin{pmatrix} \nabla_{P^{+\prime}}\Psi^{+}(\theta,P^{+}) - \lambda I_{L-1} & M_{12} \\ M_{21} - (M_{22}-\lambda)\mathbf{1}'_{L-1} & M_{22}-\lambda \end{pmatrix}.$$

Since $\det(E)=1$, we have $\det((M(\theta)-\lambda I_L)E) = \det(M(\theta)-\lambda I_L)$, and using properties of the determinant of a partitioned matrix (see, for example, 14.1 of Seber, 2007) gives

$$\det\left(M(\theta)-\lambda I_L\right) = \det\left(\nabla_{P^{+\prime}}\Psi^{+}(\theta,P^{+}) - \lambda I_{L-1}\right) \times B(\lambda),$$

where $B(\lambda) = \det(M_{22}-\lambda - [M_{21}-(M_{22}-\lambda)\mathbf{1}'_{L-1}][\nabla_{P^{+\prime}}\Psi^{+}(\theta,P^{+}) - \lambda I_{L-1}]^{-1}M_{12})$. Note that $M_{21} = \mathbf{1}'_{L-1}(I_{L-1}-M_{11})$ and $M_{22} = 1 - \mathbf{1}'_{L-1}M_{12}$ because $M(\theta)$ is a column stochastic matrix. It follows that $B(1)=0$, giving part (b).

For part (c), note that the spectral radius of $M(\theta)$ is 1 from Theorem 8.1.22 of Horn and Johnson (1985). Since $M(\theta^0)$ is irreducible and aperiodic, it follows 9.58 of Seber (2007) and Definition 8.5.0 of Horn and Johnson (1985) that only one eigenvalue of $M(\theta^0)$ has modulus one, and part (c) follows. □

7.6 Proof of Proposition 4

Let $\lambda = r\cos\theta + ir\sin\theta$ be an eigenvalue of Ψ_P. Then, the corresponding eigenvalue of Λ_P is $\lambda(\alpha) = \alpha r\cos\theta + i\alpha r\sin\theta + (1-\alpha)$. Let $f(\alpha) = |\lambda(\alpha)|^2$, then the stated result holds because $f(0)=1$ and $\nabla_\alpha f(0) = 2(r\cos\theta - 1) < 0$. □

7.7 Proof of Proposition 5

For part (a), write $\Gamma(\theta,P) - P$ as $\Gamma(\theta,P) - P = A(\theta,P)(\Psi(\theta,P)-P)$, where $A(\theta,P) \equiv (I - \Pi(\theta,P)\nabla_{P'}\Psi(\theta,P)\Pi(\theta,P))^{-1}\Pi(\theta,P) + (I-\Pi(\theta,P))$. Let $Z(\theta,P)$ denote an orthonormal basis of the column space of $\Pi(\theta,P)$, so that $Z(\theta,P)Z(\theta,P)' = \Pi(\theta,P)$ and $Z(\theta,P)'Z(\theta,P) = I_m$. Suppress (θ,P) from $\Pi(\theta,P)$, $Z(\theta,P)$, and $\nabla_{P'}\Psi(\theta,P)$. A direct calculation gives $(I - \Pi\nabla_{P'}\Psi\Pi)^{-1}\Pi = Z(I - Z'\nabla_{P'}\Psi Z)^{-1}Z'$, so we can write $A(\theta,P)$ as $A(\theta,P) = Z(I-Z'\nabla_{P'}\Psi Z)^{-1}Z' + (I-\Pi)$. The stated result follows since $A(\theta,P)$ is nonsingular because $\text{rank}[Z(I-Z'\nabla_{P'}\Psi Z)^{-1}Z'] = m$, $\text{rank}(I-\Pi) = N-m$, and $Z(I-Z'\nabla_{P'}\Psi Z)^{-1}Z'$ and $I-\Pi$ are orthogonal to each other.

For part (b), define $\Gamma_P \equiv \nabla_{P'}\Gamma(\theta^0,P^0)$ and $\Pi^0 \equiv \Pi(\theta^0,P^0)$. Define $\mathbb{P}$ with respect to $\Psi_P \equiv \nabla_{P'}\Psi(\theta^0,P^0)$. Computing $\nabla_{P'}\Gamma(\theta,P)$ and noting that $\Psi(\theta^0,P^0) = P^0$, we find $\Gamma_P = \Pi^0 + (I - \Pi^0\Psi_P\Pi^0)^{-1}\Pi^0(\Psi_P - I) + (I-\Pi^0)\Psi_P$. Observe that $\Gamma_P\Pi^0 = (I-\Pi^0)\Psi_P\Pi^0 = 0$, where the last equality follows because $\Psi_P\Pi^0 P \in \mathbb{P}$ for any $P \in \mathbb{R}^L$ by the definition of Π^0. Hence, $\Gamma_P = \Gamma_P(I-\Pi^0)$. We also have $(I-\Pi^0)\Gamma_P = (I-\Pi^0)\Psi_P$ because a direct calculation gives $(I-\Pi^0\Psi_P\Pi^0)^{-1}\Pi^0 = Z^0(I-(Z^0)'\Psi_P Z^0)^{-1}(Z^0)'$ where $Z^0 = Z(\theta^0,P^0)$, and hence $(I-\Pi^0)(I-\Pi^0\Psi_P\Pi^0)^{-1}\Pi^0 = 0$. Then, in conjunction with $\Gamma_P = \Gamma_P(I-\Pi^0)$, we obtain

$(I-\Pi^0)\Gamma_P = (I-\Pi^0)\Psi_P(I-\Pi^0)$. Since $\Gamma_P(I-\Pi^0)$ has the same eigenvalues as $(I-\Pi^0)\Gamma_P$ (see Theorem 1.3.20 of Horn and Johnson, 1985), we have $\rho(\Gamma_P) = \rho(\Gamma_P(I-\Pi^0)) = \rho((I-\Pi^0)\Gamma_P) = \rho[(I-\Pi^0)\Psi_P(I-\Pi^0)] \leq \delta^0$, where the last inequality follows from Lemma 2.10 of SK: P, Q, and F_u^* in SK correspond to our Π^0, $I-\Pi^0$, and Ψ_P. □

7.8 Proof of Proposition 6

The stated results follow from Proposition 2 of AM07 and our Lemma 1 if Assumptions 1(b)-(c) and 1(e)-(h) and Assumptions 2(b)-(c) hold when $\Psi(\theta, P)$ is replaced with $\Gamma(\theta, P)$.

We check Assumptions 2(b)-(c) first because they are used in showing the other conditions. First, note that Chu (1990, Section 4.2, in particular line 17 on page 1377) proved the following: if a matrix $A(t)$ is ℓ times continuously differentiable with respect to t, and if $X(t)$ spans the invariant subspace corresponding to a subset of eigenvalues of $A(t)$, then $X(t)$ is also ℓ times continuously differentiable with respect to t. Consequently, $\Pi(\theta, P)$ is three times continuously differentiable in $\mathcal{N}$ (we suppress "in $\mathcal{N}$" henceforth) since $\nabla_{P'}\Psi(\theta, P)$ is three times continuously differentiable from Assumption 3(b). Further, $I-\Pi(\theta, P)\nabla_{P'}\Psi(\theta, P)\Pi(\theta, P)$ is nonsingular and three times continuously differentiable from Assumptions 3(b)-(c), and hence Assumption 2(b) holds for $\Gamma(\theta, P)$. For Assumption 2(c), a direct calculation gives $\Omega^{\Gamma}_{\theta\theta} = \Psi'_\theta A(\theta^0, P^0)'\Delta_P A(\theta^0, P^0)\Psi_\theta$, where $A(\theta, P)$ is defined in the proof of Proposition 2 and shown to be nonsingular. Since $\text{rank}(\Psi_\theta) = K$ from nonsingularity of $\Omega_{\theta\theta} = \Psi'_\theta\Delta_P\Psi_\theta$, positive definiteness of $\Omega^{\Gamma}_{\theta\theta}$ follows.

We proceed to confirm Assumptions 1(b)-(c) and 1(e)-(h) hold for $\Gamma(\theta, P)$. Assumption 1(b) for $\Gamma(\theta, P)$ follows from Assumption 3(d). Assumption 1(c) holds because we have already shown that $\Gamma(\theta, P)$ is three times continuously differentiable. Assumption 1(e) holds because $\Psi(\theta, P)$ and $\Gamma(\theta, P)$ have the same fixed points by Proposition 5. As discussed in page 21 of AM07, Assumption 1(f) is implied by Assumption 3(e). Assumption 1(g) for $\tilde{\theta}^{\Gamma}_0(P)$ follows from the positive definiteness of $\Omega^{\Gamma}_{\theta\theta}$ and by the implicit function theorem applied to the first order condition for θ. Assumption 1(h) follows from Assumption 3(e). □

7.9 Proof of Proposition 7

Write the objective function as $\overline{\gamma}(\theta, P, \eta) \equiv n^{-1}\sum_{i=1}^n \ln\Gamma(\theta, P, \eta)(a_i|x_i)$, and define $\gamma(\theta, P, \eta) \equiv E\ln\Gamma(\theta, P, \eta)(a_i|x_i)$. Define $\Omega^{\Gamma}_{\theta P} \equiv E\nabla_\theta \ln\Gamma(\theta^0, P^0)(a_i|x_i)\nabla_{P'}\ln\Gamma(\theta^0, P^0)(a_i|x_i)$. For $\epsilon > 0$, define a neighborhood $\mathcal{N}_3(\epsilon) = \{(\theta, P, \eta) : \max\{||\theta-\theta^0||, ||P-P^0||, ||\eta-\theta^0||\} < \epsilon\}$. Then, there exists $\epsilon_1 > 0$ such that (i) $\mathcal{N}(\epsilon_1) \subset \mathcal{N}$, (ii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'}\gamma(\theta, P, \eta)^{-1}|| < \infty$, and (iii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)} ||\nabla^3\gamma(\theta, P, \eta)|| < \infty$ because $\Gamma(\theta^0, P^0, \theta^0)(a|x) = P^0(a|x) > 0$, $\Gamma(\theta, P, \eta)$ is three times continuously differentiable (see the proof of Proposition 6), and $\nabla_{\theta\theta'}\gamma(\theta^0, P^0, \theta^0) = \nabla_{\theta\theta'}\gamma(\theta^0, P^0)$ is nonsingular.

First, we assume $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ and derive the stated representation of $\tilde{\theta}_j - \hat{\theta}$ and $\tilde{P}_j - \hat{P}$. We later show $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ wpa1 if c is taken sufficiently small. Henceforth,

we suppress the subscript RPM from $\hat{\theta}_{RPM}$ and $\hat{P}_{RPM}$. Expanding the first order condition $\nabla_\theta \overline{\gamma}(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = 0$ around $(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ gives

$$0 = \nabla_\theta \overline{\gamma}(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) + \nabla_{\theta\theta'} \overline{\gamma}(\bar{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})(\tilde{\theta}_j - \hat{\theta}), \tag{14}$$

where $\bar{\theta} \in [\tilde{\theta}_j, \hat{\theta}]$. Writing $\bar{\theta} = \bar{\theta}(\tilde{\theta}_j)$, we obtain $\sup_{(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} \overline{\gamma}(\bar{\theta}(\tilde{\theta}_j), \tilde{P}_{j-1}, \tilde{\theta}_{j-1})^{-1}|| = O_p(1)$ because (i) $||\bar{\theta}(\tilde{\theta}_j) - \theta^0|| < \epsilon_1$ wpa1 since $||\tilde{\theta}_j - \theta^0|| < \epsilon_1$ and $\hat{\theta}$ is consistent, and (ii) $\sup_{(\theta,P,\eta) \in \mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} \overline{\gamma}(\theta, P, \eta)^{-1}|| = O_p(1)$ since $\sup_{(\theta,P,\eta) \in \mathcal{N}_3(\epsilon_1)} ||\nabla_{\theta\theta'} \gamma(\theta, P, \eta)^{-1}|| < \infty$ and $\sup_{(\theta,P,\eta) \in \mathcal{N}_3(\epsilon_1)} ||\nabla^2 \overline{\gamma}(\theta, P, \eta) - \nabla^2 \gamma(\theta, P, \eta)|| = o_p(1)$. Therefore, the stated representation of $\tilde{\theta}_j - \hat{\theta}$ follows if we show

$$\nabla_\theta \overline{\gamma}(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = -\Omega^\Gamma_{\theta P}(\tilde{P}_{j-1} - \hat{P}) + r_{nj}, \tag{15}$$

where r_{nj} denotes a generic remainder term that is $O_p(n^{-1/2}||\tilde{\theta}_{j-1} - \hat{\theta}|| + ||\tilde{\theta}_{j-1} - \hat{\theta}||^2 + n^{-1/2}||\tilde{P}_{j-1} - \hat{P}|| + ||\tilde{P}_{j-1} - \hat{P}||^2)$ uniformly in $(\tilde{\theta}_{j-1}, \tilde{P}_{j-1}) \in \mathcal{N}(\epsilon_1)$.

We proceed to show (15). Expanding $\nabla_\theta \overline{\gamma}(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ twice around $(\hat{\theta}, \hat{P}, \hat{\theta})$ gives $\nabla_\theta \overline{\gamma}(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = \nabla_\theta \overline{\gamma}(\hat{\theta}, \hat{P}, \hat{\theta}) + \nabla_{\theta P'} \overline{\gamma}(\hat{\theta}, \hat{P}, \hat{\theta})(\tilde{P}_{j-1} - \hat{P}) + \nabla_{\theta\eta'} \overline{\gamma}(\hat{\theta}, \hat{P}, \hat{\theta})(\tilde{\theta}_{j-1} - \hat{\theta}) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}||^2 + ||\tilde{P}_{j-1} - \hat{P}||^2)$. For the first term on the right, the RPM estimator satisfies $\nabla_\theta \overline{\gamma}(\hat{\theta}, \hat{P}, \hat{\theta}) = 0$ wpa1 because $\nabla_{\theta'} \overline{\gamma}(\hat{\theta}, \hat{P}) = 0$ from the first order condition, and Proposition 5(a) implies $\Psi(\hat{\theta}, \hat{P}) = \hat{P}$ wpa1 and hence $\nabla_{\theta'} \Gamma(\hat{\theta}, \hat{P}, \hat{\theta}) = \nabla_{\theta'} \Gamma(\hat{\theta}, \hat{P})$ wpa1. For the second and third terms on the right, we have $E \nabla_{\theta P'} \ln \Gamma(\theta^0, P^0, \theta^0)(a_i|x_i) = -\Omega^\Gamma_{\theta P}$ and $E \nabla_{\theta\eta'} \ln \Gamma(\theta^0, P^0, \theta^0)(a_i|x_i) = 0$ by the information matrix equality because $\Gamma(\theta^0, P^0, \theta^0) = \Gamma(\theta^0, P^0)$, $\nabla_{\theta'} \Gamma(\theta^0, P^0, \theta^0) = \nabla_{\theta'} \Gamma(\theta^0, P^0)$, $\nabla_{P'} \Gamma(\theta^0, P^0, \theta^0) = \nabla_{P'} \Gamma(\theta^0, P^0)$, and $\nabla_{\eta'} \Gamma(\theta^0, P^0, \theta^0) = 0$ from $P^0 = \Psi(\theta^0, P^0)$. Therefore, (15) follows from the root-n consistency of $(\hat{\theta}, \hat{P})$.

For the representation of $\tilde{P}_j - \hat{P}$, first we have

$$\tilde{P}_j = \hat{P} + \Gamma_\theta(\tilde{\theta}_j - \hat{\theta}) + \Gamma_P(\tilde{P}_{j-1} - \hat{P}) + r_{nj}, \tag{16}$$

by expanding $\tilde{P}_j = \Gamma(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_j)$ around $(\hat{\theta}, \hat{P}, \hat{\theta})$ and using $\Gamma(\hat{\theta}, \hat{P}, \hat{\theta}) = \hat{P}$. Next, refine (14) as $0 = \nabla_\theta \overline{\gamma}(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) - \Omega^\Gamma_{\theta\theta}(\tilde{\theta}_j - \hat{\theta}) + r_{nj}$ by expanding $\nabla_{\theta\theta'} \overline{\gamma}(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1})$ in (14) around $(\hat{\theta}, \hat{P}, \hat{\theta})$ to write it as $\nabla_{\theta\theta'} \overline{\gamma}(\hat{\theta}, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) = -\Omega^\Gamma_{\theta\theta} + O_p(n^{-1/2}) + O_p(||\tilde{\theta}_{j-1} - \hat{\theta}||) + O_p(||\tilde{P}_{j-1} - \hat{P}||)$ and using the bound of $\tilde{\theta}_j - \hat{\theta}$ obtained above. Substituting this into (15) gives

$$\tilde{\theta}_j - \hat{\theta} = -(\Omega^\Gamma_{\theta\theta})^{-1} \Omega^\Gamma_{\theta P}(\tilde{P}_{j-1} - \hat{P}) + r_{nj}. \tag{17}$$

The stated result follows from substituting this into (16) in conjunction with $(\Omega^\Gamma_{\theta\theta})^{-1} \Omega^\Gamma_{\theta P} = (\Gamma'_\theta \Delta_P \Gamma_\theta)^{-1} \Gamma'_\theta \Delta_P \Gamma_P$.

It remains to show $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ wpa1 if c is taken sufficiently small. We first

show that

$$\sup_{(\theta,\eta,P)\in\bar{\Theta}_j\times\mathcal{N}} |\overline{\gamma}(\theta,P,\eta)-\gamma(\theta,P,\eta)| = o_p(1), \quad \gamma(\theta,P,\eta) \text{ is continuous in } (\theta,\eta,P)\in\bar{\Theta}_j\times\mathcal{N}. \tag{18}$$

Take $\mathcal{N}$ sufficiently small, then it follows from the consistency of $(\tilde{\theta}_{j-1},\tilde{P}_{j-1})$ and the continuity of $\Gamma(\theta,P,\eta)$ that $\Gamma(\theta,P,\eta)(a|x)\in[\xi/2,1-\xi/2]$ for all $(a,x)\in A\times X$ and $(\theta,P,\eta)\in\bar{\Theta}_j\times\mathcal{N}$ wpa1. Observe that (i) $\bar{\Theta}_j\times\mathcal{N}$ is compact because it is an intersection of the compact set Θ and $|A||X|$ closed sets, (ii) $\ln\Gamma(\theta,P,\eta)$ is continuous in $(\theta,P,\eta)\in\bar{\Theta}_j\times\mathcal{N}$, and (iii) $E\sup_{(\theta,P,\eta)\in\bar{\Theta}_j\times\mathcal{N}}|\ln\Gamma(\theta,P,\eta)(a_i|x_i)|\leq|\ln(\xi/2)|+|\ln(1-\xi/2)|<\infty$ because of the way we choose $\mathcal{N}$. Therefore, (18) follows from Lemma 2.4 of Newey and McFadden (1994).

Finally, we show $(\tilde{\theta}_j,\tilde{P}_{j-1},\tilde{\theta}_{j-1})\in\mathcal{N}_3(\epsilon_1)$ wpa1 under (18) by applying the argument in the proof of Lemma 1. Define $\Delta=\gamma(\theta^0,P^0,\theta^0)-\sup_{\theta\in\mathcal{N}_\theta(\epsilon_1)^c\cap\Theta}\gamma(\theta,P^0,\theta^0)>0$, where the last inequality follows from the information inequality because $\gamma(\theta,P^0,\theta^0)$ is uniquely maximized at θ^0 and $\mathcal{N}_\theta(\epsilon_1)^c\cap\Theta$ is compact. It follows that $\Pr(\tilde{\theta}_j\notin\mathcal{N}_\theta(\epsilon_1))\leq\Pr(\gamma(\theta^0,P^0,\theta^0)-\gamma(\tilde{\theta}_j,P^0,\theta^0)\geq\Delta)$. Proceeding as in the proof of Lemma 1, we find that, if c is taken sufficiently small, then $\gamma(\theta^0,P^0,\theta^0)-\gamma(\tilde{\theta}_j,P^0,\theta^0)\leq\Delta/2+o_p(1)$ and hence $\Pr((\tilde{\theta}_j,\tilde{P}_{j-1},\tilde{\theta}_{j-1})\notin\mathcal{N}(\epsilon_1))\to 0$. □

7.10 Proof of Proposition 8

The proof closely follows the proof of Lemma 2. We suppress the subscript RPM from $\hat{\theta}_{RPM}$ and $\hat{P}_{RPM}$. Let $\tilde{\zeta}_j=(\tilde{\theta}_j',\tilde{P}_j')'$ and $\hat{\zeta}=(\hat{\theta}',\hat{P}')'$. Let $b>0$ be a constant such that $\rho(M_{\Gamma_\theta}\Gamma_P)+2b<1$. Define

$$D=\begin{pmatrix}0 & -(\Omega^\Gamma_{\theta\theta})^{-1}\Omega^\Gamma_{\theta P}\\ 0 & M_{\Gamma_\theta}\Gamma_P\end{pmatrix}. \tag{19}$$

Note that $\rho(D)=\rho(M_{\Gamma_\theta}\Gamma_P)$ and there exists a matrix norm $||\cdot||_\alpha$ such that $\|D\|_\alpha\leq\rho(D)+b=\rho(M_{\Gamma_\theta}\Gamma_P)+b$. We define the vector norm for $x\in\mathbb{R}^{k+L}$ as $||x||_\beta=||[x\ 0\dots0]||_\alpha$, then $||Ax||_\beta\leq||A||_\alpha||x||_\beta$ for any matrix A.

From the representation of $\tilde{P}_j-\hat{P}$ and $\tilde{\theta}_j-\hat{\theta}$ in Proposition 7 and (17), and the equivalence of vector norms in $\mathbb{R}^{k+L}$, there exists $c>0$ such that $\tilde{\zeta}_j-\hat{\zeta}=D(\tilde{\zeta}_{j-1}-\hat{\zeta})+O_p(n^{-1/2}||\tilde{\zeta}_{j-1}-\hat{\zeta}||_\beta+||\tilde{\zeta}_{j-1}-\hat{\zeta}||_\beta^2))$ holds uniformly in $\tilde{\zeta}_{j-1}\in\{\zeta:||\zeta-\zeta^0||_\beta<c\}$. The stated result then follows from repeating the proof of Lemma 2. □

7.11 Proof of Proposition 9

Part (a) follows from Proposition 2 of AM07 if Assumptions 1(b)-(c) and 1(e)-(h) and Assumptions 2(b)-(c) hold when $\Psi(\theta,P)$ is replaced with $\Lambda^q(\theta,P)$. Similar to the proof of Proposition 7, we check Assumptions 2(b)-(c) first. Assumption 2(b) holds for $\Lambda^q(\theta,P)$ because $\Psi(\theta,P)$ is three times continuously differentiable in $\mathcal{N}$ from Assumption 4(b). For Assumption 2(c), a direct calculation gives $\Omega^q_{\theta\theta}=(\nabla_{\theta'}\Lambda^q(\theta^0,P^0))'\Delta_P\nabla_{\theta'}\Lambda^q(\theta^0,P^0)=\Lambda_\theta'(I-(\Lambda_P)^q)'(I-$

$\Lambda_P')^{-1}\Delta_P(I-\Lambda_P)^{-1}(I-(\Lambda_P)^q)\Lambda_\theta = \Psi_\theta'(I-(\alpha\Psi_P+(1-\alpha)I)^q)'(I-\Psi_P')^{-1}\Delta_P(I-\Psi_P)^{-1}(I-(\alpha\Psi_P+(1-\alpha)I)^q)\Psi_\theta$, where the second equality follows from $\nabla_{\theta'}\Lambda^q(\theta^0,P^0) = (\sum_{j=0}^{q-1}(\Lambda_P)^j)\Lambda_\theta = (I-\Lambda_P)^{-1}(I-(\Lambda_P)^q)\Lambda_\theta$, and the third equality follows from $\Lambda_\theta = \alpha\Psi_\theta$ and $\Lambda_P = \alpha\Psi_P+(1-\alpha)I$. Since $\text{rank}(\Psi_\theta) = K$ from nonsingularity of $\Omega_{\theta\theta} = \Psi_\theta'\Delta_P\Psi_\theta$, positive definiteness of $\Omega^q_{\theta\theta}$ follows from Assumption 4(d).

The proof of part (a) is completed by confirming that Assumptions 1(b)-(c) and 1(e)-(h) hold for $\Lambda^q(\theta,P)$. Assumptions 1(b)-(c) hold for $\Lambda^q(\theta,P)$ because Assumptions 1(b)-(c) hold for $\Psi(\theta,P)$. Assumption 1(e) for $\Lambda^q(\theta,P)$ follows from Assumption 4(c). As discussed in page 21 of AM07, Assumption 1(f) for $\Lambda^q(\theta,P)$ is implied by Assumption 4(e). Assumption 1(g) for $\tilde\theta_0^q(P)$ follows from the positive definiteness of $\Omega^q_{\theta\theta}$ and applying the implicit function theorem to the first order condition for θ. Assumption 1(h) follows from Assumption 4(e). This completes the proof of part (a).

We proceed to prove part (b). Define the objective function and its limit as $Q_n^q(\theta,P,\eta) \equiv n^{-1}\sum_{i=1}^n \ln\Lambda^q(\theta,P,\eta)(a_i|x_i)$ and $Q^q(\theta,P,\eta) \equiv E\ln\Lambda^q(\theta,P,\eta)(a_i|x_i)$. For $\epsilon>0$, define a neighborhood $\mathcal{N}_3(\epsilon) = \{(\theta,P,\eta) : \max\{||\theta-\theta^0||,||P-P^0||,||\eta-\theta^0||\} < \epsilon\}$. Then, there exists $\epsilon_1 > 0$ such that (i) $\mathcal{N}(\epsilon_1)\subset\mathcal{N}$, (ii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)}||\nabla_{\theta\theta'}Q^q(\theta,P,\eta)^{-1}|| < \infty$, and (iii) $\sup_{(\theta,P,\eta)\in\mathcal{N}_3(\epsilon_1)}||\nabla^3 Q^q(\theta,P,\eta)|| < \infty$ because $\Lambda^q(\theta^0,P^0,\theta^0)(a|x) = P^0(a|x) > 0$, $\Lambda^q(\theta,P,\eta)$ is three times continuously differentiable, and $\nabla_{\theta\theta'}Q^q(\theta^0,P^0,\theta^0) = \nabla_{\theta\theta'}Q^q(\theta^0,P^0)$ is nonsingular.

First, we assume $(\tilde\theta_j,\tilde P_{j-1},\tilde\theta_{j-1})\in\mathcal{N}_3(\epsilon_1)$ and derive the stated representation of $\tilde\theta_j-\hat\theta$ and $\tilde P_j-\hat P$. We later show $(\tilde\theta_j,\tilde P_{j-1},\tilde\theta_{j-1})\in\mathcal{N}_3(\epsilon_1)$ wpa1 if c is taken sufficiently small. Henceforth, we suppress the subscript qNPL from $\hat\theta_{qNPL}$ and $\hat P_{qNPL}$. The proof is similar to the proof of the updating formula of Proposition 7. For the representation of $\tilde\theta_j-\hat\theta$, expanding the first order condition $0 = \nabla_\theta Q_n^q(\tilde\theta_j,\tilde P_{j-1},\tilde\theta_{j-1})$ around $(\hat\theta,\tilde P_{j-1},\tilde\theta_{j-1})$ gives $0 = \nabla_\theta Q_n^q(\hat\theta,\tilde P_{j-1},\tilde\theta_{j-1}) + \nabla_{\theta\theta'}Q_n^q(\bar\theta(\tilde\theta_j),\tilde P_{j-1},\tilde\theta_{j-1})(\tilde\theta_j-\hat\theta)$, which corresponds to (14) in the proof of Proposition 7. Proceeding as in the proof of Proposition 7, we obtain $\sup_{(\tilde\theta_j,\tilde P_{j-1},\tilde\theta_{j-1})\in\mathcal{N}_3(\epsilon_1)}||\nabla_{\theta\theta'}Q_n^q(\bar\theta(\tilde\theta_j),\tilde P_{j-1},\tilde\theta_{j-1})^{-1}|| = O_p(1)$. Therefore, the stated representation of $\tilde\theta_j-\hat\theta$ follows if we show $\nabla_\theta Q_n^q(\hat\theta,\tilde P_{j-1},\tilde\theta_{j-1}) = -\Omega^q_{\theta P}(\tilde P_{j-1}-\hat P)+r_{nj}$, where r_{nj} denotes a remainder term of $O_p(n^{-1/2}||\tilde\theta_{j-1}-\hat\theta|| + ||\tilde\theta_{j-1}-\hat\theta||^2 + n^{-1/2}||\tilde P_{j-1}-\hat P|| + ||\tilde P_{j-1}-\hat P||^2)$ uniformly in $(\tilde\theta_{j-1},\tilde P_{j-1})\in\mathcal{N}(\epsilon_1)$. This representation corresponds to (15) in the proof of Proposition 7 and follows from the same argument. Namely, expanding $\nabla_\theta Q_n^q(\hat\theta,\tilde P_{j-1},\tilde\theta_{j-1})$ twice around $(\hat\theta,\hat P,\hat\theta)$ and noting that (i) the q-NPL estimator satisfies $\nabla_\theta Q_n^q(\hat\theta,\hat P,\hat\theta)=0$, (ii) $\Lambda^q(\theta^0,P^0,\theta^0) = \Lambda^q(\theta^0,P^0)$, $\nabla_{\theta'}\Lambda^q(\theta^0,P^0,\theta^0) = \nabla_{\theta'}\Lambda^q(\theta^0,P^0)$, $\nabla_{P'}\Lambda^q(\theta^0,P^0,\theta^0) = \nabla_{P'}\Lambda^q(\theta^0,P^0)$, and $\nabla_{\eta'}\Lambda^q(\theta^0,P^0,\theta^0)=0$, and using the information matrix equality and the root-n consistency of $(\hat\theta,\hat P)$ gives the required result.

The proof of the representation of $\tilde P_j-\hat P$ follows from the proof of Proposition 7, because (i) $\tilde P_j = \hat P + \Lambda^q_\theta(\tilde\theta_j-\hat\theta) + \Lambda^q_P(\tilde P_{j-1}-\hat P)+r_{nj}$, which corresponds to (16) in the proof of Proposition 7, from expanding $\Lambda^q(\tilde\theta_j,\tilde P_{j-1})$ twice around $(\hat\theta,\hat P)$ and using $\hat P = \Lambda^q(\hat\theta,\hat P)$, (ii) $\nabla_{\theta\theta'}Q_n^q(\hat\theta,\tilde P_{j-1},\tilde\theta_{j-1})(\tilde\theta_j-\hat\theta) = -\Omega^q_{\theta\theta}(\tilde\theta_j-\hat\theta)+r_{nj}$ from expanding $\nabla_{\theta\theta'}Q_n^q(\hat\theta,\tilde P_{j-1},\tilde\theta_{j-1})$ around $(\hat\theta,\hat P,\hat\theta)$ and using the bound of $\tilde\theta_j-\hat\theta$ obtained above, and (iii) $(\Omega^q_{\theta\theta})^{-1}\Omega^q_{\theta P} = ((\Lambda^q_\theta)'\Delta_P\Lambda^q_\theta)^{-1}(\Lambda^q_\theta)'\Delta_P\Lambda^q_P$.

The proof of part (b) is completed by showing $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ wpa1 if c is taken sufficiently small. First, observe that (18) in the proof of Proposition 7 holds with $\overline{\gamma}(\theta, P, \eta)$ and $\gamma(\theta, P, \eta)$ replacing $Q_n^q(\theta, P, \eta)$ and $Q^q(\theta, P, \eta)$ if we take $\mathcal{N}$ sufficiently small. Therefore, $(\tilde{\theta}_j, \tilde{P}_{j-1}, \tilde{\theta}_{j-1}) \in \mathcal{N}_3(\epsilon_1)$ wpa1 follows from repeating the argument in the last paragraph of the proof of Proposition 7 if we show that θ^0 uniquely maximizes $Q^q(\theta, P^0, \theta^0)$. Note that

$$\begin{aligned} Q^q(\theta, P^0, \theta^0) - Q^q(\theta^0, P^0, \theta^0) &= E \ln(\nabla_{\theta'} \Lambda^q(\theta^0, P^0)(\theta - \theta^0) + P^0)(a_i|x_i) - E \ln P^0(a_i|x_i) \\ &= E \ln \left(\frac{\nabla_{\theta'} \Lambda^q(\theta^0, P^0)(a_i|x_i)(\theta - \theta^0)}{P^0(a_i|x_i)} + 1 \right). \end{aligned} \tag{20}$$

Recall that $\ln(y+1) \leq y$ for all $y > -1$ where the inequality is strict if $y \neq 0$. Since $\text{rank}(\nabla_{\theta'} \Lambda^q(\theta^0, P^0)) = K$ from the positive definiteness of $\Omega^q_{\theta\theta}$, it follows that $\nabla_{\theta'} \Lambda^q(\theta^0, P^0)\nu \neq 0$ for any K-vector $\nu \neq 0$. Therefore, $\nabla_{\theta'} \Lambda^q(\theta^0, P^0)(a_i|x_i)(\theta - \theta^0) \neq 0$ for at least one (a_i, x_i) for all $\theta \neq \theta^0$. Consequently, the right hand side of (20) is strictly smaller than $E[\nabla_{\theta'} \Lambda^q(\theta^0, P^0)(a_i|x_i)(\theta - \theta^0)/P^0(a_i|x_i)]$ for all $\theta \neq \theta^0$. Because $E[\nabla_{\theta'} \Lambda^q(\theta^0, P^0)(a_i|x_i)/P^0(a_i|x_i)] = 0$, we have $Q^q(\theta, P^0, \theta^0) - Q^q(\theta^0, P^0, \theta^0) < 0$ for all $\theta \neq \theta^0$. Therefore, θ^0 uniquely maximizes $Q(\theta, P^0, \theta^0)$, and we complete the proof of part (b).

We prove part (c). From the proof of part (a) in conjunction with the relation $\Lambda_P = \alpha \Psi_P + (1-\alpha) I$, we may write $\Omega^q_{\theta\theta}$ as $\Omega^q_{\theta\theta} = \Psi'_\theta (I - (\Lambda_P)^q)' (I - \Psi'_P)^{-1} \Delta_P (I - \Psi_P)^{-1} (I - (\Lambda_P)^q) \Psi_\theta$. Similarly, using the relation $\nabla_{P'} \Lambda^q(\theta^0, P^0) = (\Lambda_P)^q$, we obtain $\Omega^q_{\theta P} = \Lambda'_\theta (I - (\Lambda_P)^q)' (I - \Lambda'_P)^{-1} \Delta_P (\Lambda_P)^q$. Therefore, if $\rho(\Lambda_P) < 1$, then $\Omega^q_{\theta\theta} \to \Psi'_\theta (I - \Psi'_P)^{-1} \Delta_P (I - \Psi_P)^{-1} \Psi_\theta$ and $\Omega^q_{\theta P} \to 0$ as $q \to \infty$, and it follows that $V_{qNPL} \to (\Psi'_\theta (I - \Psi'_P)^{-1} \Delta_P (I - \Psi_P)^{-1} \Psi_\theta)^{-1}$ as $q \to \infty$. This limit is the same as $V_{MLE} = (E[\nabla_\theta \ln P(\theta^0)(a_i|x_i) \nabla_{\theta'} \ln P(\theta^0)(a_i|x_i)])^{-1}$, where $P(\theta) \equiv \arg\max_{P \in \mathcal{M}_\theta} E \ln P(a_i|x_i)$ with $\mathcal{M}_\theta \equiv \{P \in B_P : P = \Psi(\theta, P)\}$, because $\nabla_{\theta'} P(\theta) = (I - \nabla_{P'} \Psi(\theta, P(\theta)))^{-1} \nabla_{\theta'} \Psi(\theta, P(\theta))$ holds in a neighborhood of $\theta = \theta^0$.

We omit the proof of part (d) because it is identical to the proof of Proposition 8 except that $\hat{\theta}_{RPM}$, $\hat{P}_{RPM}$, $(\Omega^\Gamma_{\theta\theta})^{-1} \Omega^\Gamma_{\theta P}$, and $M_{\Gamma_\theta} \Gamma_P$ are replaced with $\hat{\theta}_{qNPL}$, $\hat{P}_{qNPL}$, $(\Omega^q_{\theta\theta})^{-1} \Omega^q_{\theta P}$, and $M_{\Lambda^q_\theta} \Lambda^q_P$, respectively. □