Supplementary Appendix to "Pseudo-likelihood Estimation and Bootstrap Inference for Structural Discrete Markov Decision Models"

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

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

June 11, 2008

This supplementary appendix contains the following technical details omitted from the main paper due to space constraints: (i) the convergence rate of the one-step NPL algorithm, (ii) the proof of Lemma 3, and (iii) the proof of Lemmas 7-10 in Appendix B.

1 One-step NPL Algorithm

Let $L_N(P,\alpha,\theta_f) = N^{-1}\sum_{i=1}^{N} \ln \Psi(P,\alpha,\theta_f)(a_i|x_i)$ be the NPL objective function. Suppose that an initial consistent estimator of α is available. The one-step NPL algorithm, with its estimator denoted by $(\tilde{\alpha}_k^{PL}, \tilde{P}_k^{PL})$, is defined recursively as:

Step 1: Given $(\tilde{P}_{j-1}^{PL}, \tilde{\alpha}_{j-1}^{PL}, \hat{\theta}_f)$, update α by $\tilde{\alpha}_j^{PL} = \tilde{\alpha}_{j-1}^{PL} - (Q_{N,j-1})^{-1}\frac{\partial}{\partial\alpha'}L_N(\tilde{P}_{j-1}^{PL}, \tilde{\alpha}_{j-1}^{PL}, \hat{\theta}_f)$, where $Q_{N,j-1} = Q_N(\tilde{P}_{j-1}^{PL}, \tilde{\alpha}_{j-1}^{PL}, \hat{\theta}_f)$.

Step 2: Update P using $\tilde{\alpha}_j^{PL}$ by $\tilde{P}_j^{PL} = \Psi(\tilde{P}_{j-1}^{PL}, \tilde{\alpha}_j^{PL}, \hat{\theta}_f)$.

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

The following proposition establishes that the one-step NPL algorithm achieves a similar rate of convergence to the original NPL algorithm.

Proposition A.1 *Suppose the assumptions of Proposition 2 hold and the initial estimates $(\tilde{\alpha}_0^{PL}, \tilde{P}_0^{PL})$ are consistent. Then, for $k = 1, 2, \ldots$*

$$\begin{aligned}
\tilde{\alpha}_k^{PL} - \hat{\alpha} &= O_p(||\tilde{\alpha}_{k-1}^{PL} - \hat{\alpha}||^2 + N^{-1/2}||\tilde{P}_{k-1}^{PL} - \hat{P}|| + ||\tilde{P}_{k-1}^{PL} - \hat{P}||^2) \\
&\quad [+O_p(N^{-1/2}||\hat{\alpha} - \tilde{\alpha}_{k-1}^{PL}||) \quad \text{for OPG }], \\
\tilde{P}_k^{PL} - \hat{P} &= O_p(||\tilde{\alpha}_k^{PL} - \hat{\alpha}||).
\end{aligned}$$

Proof of Proposition A.1 We prove the result for only the NR and OPG methods. The proof for the default NR and line-search NR is essentially the same except for showing $\Pr(Q_N^D \neq Q_N^{NR}) \to 0$ and $\Pr(Q_N^{LS} \neq Q_N^{NR}) \to 0$; see the proof of Lemma 7.1 of Andrews (2005) (A05 hereafter). We suppress the superscript PL from $\tilde{\alpha}_j^{PL}$ and $\tilde{P}_j^{PL}$, and we suppress $\hat{\theta}_f$ from $\overline{\psi}_\alpha(P,\alpha,\hat{\theta}_f)$ and $Q_N(P,\alpha,\hat{\theta}_f)$ when it does not lead to confusion.

Recall the MLE satisfies the first order condition $\overline{\psi}_\alpha(\hat{P},\hat{\alpha}) = 0$. Applying the generalized Taylor's theorem to $\overline{\psi}_\alpha(\hat{P},\hat{\alpha}) - \overline{\psi}_\alpha(\tilde{P}_{j-1},\tilde{\alpha}_{j-1})$ gives

$$\begin{aligned}
0 &= \overline{\psi}_\alpha(\tilde{P}_{j-1},\tilde{\alpha}_{j-1}) + D_\alpha\overline{\psi}_\alpha(\tilde{P}_{j-1},\tilde{\alpha}_{j-1})(\hat{\alpha}-\tilde{\alpha}_{j-1}) \\
&\quad + D_P\overline{\psi}_\alpha(\tilde{P}_{j-1},\tilde{\alpha}_{j-1})(\hat{P}-\tilde{P}_{j-1}) + R_{N,j} \\
&= \overline{\psi}_\alpha(\tilde{P}_{j-1},\tilde{\alpha}_{j-1}) + Q_N(\tilde{P}_{j-1},\tilde{\alpha}_{j-1})(\tilde{\alpha}_j-\tilde{\alpha}_{j-1}) + Q_N(\tilde{P}_{j-1},\tilde{\alpha}_{j-1})(\hat{\alpha}-\tilde{\alpha}_j) \\
&\quad + \left[D_\alpha\overline{\psi}_\alpha(\tilde{P}_{j-1},\tilde{\alpha}_{j-1}) - Q_N(\tilde{P}_{j-1},\tilde{\alpha}_{j-1})\right](\hat{\alpha}-\tilde{\alpha}_{j-1}) \\
&\quad + D_P\overline{\psi}_\alpha(\tilde{P}_{j-1},\tilde{\alpha}_{j-1})(\hat{P}-\tilde{P}_{j-1}) + R_{N,j}, \qquad \text{(A-1)}
\end{aligned}$$

where $R_{N,j} = O_p(||\hat{P}-\tilde{P}_{j-1}||^2 + ||\hat{\alpha}-\tilde{\alpha}_{j-1}||^2)$ from Lemma 7(b). The first two terms on the right of (A-1) cancel out. For the fourth term on the right of (A-1), the term inside the bracket is zero in the NR and $O_p(||\hat{P}-\tilde{P}_{j-1}|| + ||\hat{\alpha}-\tilde{\alpha}_{j-1}|| + N^{-1/2})$ in the OPG from Lemma 7(d), (e) and the information matrix equality. For the fifth term on the right of (A-1), it follows from the generalized Taylor's theorem, Lemma 7(c), and $\hat{P}-P^0, \hat{\theta}-\theta^0 = O_p(N^{-1/2})$ that $D_P\overline{\psi}_\alpha(\tilde{P}_{j-1},\tilde{\alpha}_{j-1},\hat{\theta}_f) = O_p(||\tilde{P}_{j-1}-\hat{P}||) + O_p(||\tilde{\alpha}_{j-1}-\hat{\alpha}||) + O_p(N^{-1/2})$. Therefore, $Q_N(\tilde{P}_{j-1},\tilde{\alpha}_{j-1})(\hat{\alpha}-\tilde{\alpha}_j) = O_p(N^{-1/2}||\hat{P}-\tilde{P}_{j-1}||) + O_p(||\hat{\alpha}-\tilde{\alpha}_{j-1}||^2 + ||\hat{P}-\tilde{P}_{j-1}||^2)$ $[+O_p(N^{-1/2}||\hat{\alpha}-\tilde{\alpha}_{j-1}||)$ for OPG]. The stated bound of $\tilde{\alpha}_j - \hat{\alpha}$ follows from $Q_N(\tilde{P}_{j-1},\tilde{\alpha}_{j-1}) \to_p E(\partial^2/\partial\alpha\partial\alpha')\ln\Psi(P^0,\theta^0)$, which is negative definite.

We complete the proof by showing the bound of $\tilde{P}_j - \hat{P}$. Similarly to the proof of Proposition 2, expanding $\tilde{P}_j = \Psi(\tilde{P}_{j-1},\tilde{\alpha}_j)$ around $(\hat{P},\hat{\alpha})$ and applying $D_P\Psi(\hat{P},\hat{\alpha}) = 0$ and Assumption 4(g) gives $\tilde{P}_j = \hat{P} + O_p(||\tilde{\alpha}_j-\hat{\alpha}|| + ||\tilde{P}_{j-1}-\hat{P}||^2) = \hat{P} + O_p(||\tilde{\alpha}_j-\hat{\alpha}||)$. The required result follows by induction. □

2 Proof of Lemma 3

We drop the superscript PL and MPL from $\tilde{\alpha}_k$ and $\tilde{P}_k$. We show that, if $\tilde{\alpha}_0 = \alpha^0$ and $\tilde{P}_0 = P^0$, then for $k = 0,1,\ldots$ (this corresponds to (A.9) of A05)

$$\sup_{\theta^0\in\Theta_1} \Pr_{\theta^0}(||\tilde{\alpha}_k-\hat{\alpha}|| > \mu_{N,k}) = o(N^{-c}), \qquad \sup_{\theta^0\in\Theta_1} \Pr_{\theta^0}\left(||\tilde{P}_k-\hat{P}|| > \mu_{N,k}\right) = o(N^{-c}), \qquad \text{(A-2)}$$

$$\sup_{\theta^0\in\Theta_1} \Pr_{\theta^0}\left(|T_{N,k}(\theta_r^0) - T_N(\theta_r^0)| > N^{-1/2}\mu_{N,k}\right) = o(N^{-c}), \qquad \text{(A-3)}$$

$$\sup_{\theta^0\in\Theta_1} \Pr_{\theta^0}\left(|\mathcal{W}_{N,k}(\theta^0) - \mathcal{W}_N(\theta^0)| > N^{-1/2}\mu_{N,k}\right) = o(N^{-c}). \qquad \text{(A-4)}$$

Then, as in the proof of Theorem 7.1 of A05 (p. 203), the stated result follows from applying Lemma A.1 of A05 three times, because the condition on $\hat{\theta}$ (corresponding to $\hat{\theta}_N$ in A05) in Lemma A.1 of A05 is satisfied by our Lemma 9.

First, using an induction argument, we prove the result for the one-step NPL algorithm. Let $\mu_{N,k} = N^{-(k+1)/2} \ln^{k+1} N$. For $k = 0$, (A-2) holds from Lemma 9 and $\sup_{\theta \in \Theta} ||(\partial/\partial\theta) P_\theta|| < \infty$. Suppose (A-2) holds for $k = j - 1 \geq 0$. Then, from (A-1) in the proof of Proposition A.1, we have

$$\begin{aligned}\tilde{\alpha}_j - \hat{\alpha} \;=\; & Q_N(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})^{-1} \left[D_\alpha \overline{\psi}_\alpha(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1}) - Q_N(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1}) \right] (\hat{\alpha} - \tilde{\alpha}_{j-1}) \\ & + Q_N(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})^{-1} D_P \overline{\psi}_\alpha(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})(\hat{P} - \tilde{P}_{j-1}) + Q_N(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})^{-1} R_{N,j} \quad \text{(A-5)}\end{aligned}$$

where $||R_{N,j}|| \leq (\sup_{(P,\alpha,\theta_f)} ||D^2 \overline{\psi}_\alpha(P, \alpha, \theta_f)||)(||\hat{\alpha} - \tilde{\alpha}_{j-1}||^2 + ||\hat{P} - \tilde{P}_{j-1}||^2)$.

We obtain $||D_P \overline{\psi}_\alpha(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})|| \leq \xi_{N,j}(N^{-1/2} \ln N + ||\tilde{P}_{j-1} - \hat{P}|| + ||\tilde{\alpha}_{j-1} - \hat{\alpha}||)$ with $\sup_{\theta^0 \in \Theta_1} \Pr_{\theta^0}(||\xi_{N,j}|| > K) = o(N^{-c})$ for some $K < \infty$, by expanding $D_P \overline{\psi}_\alpha(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1}) = D_P \overline{\psi}_\alpha(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1}, \hat{\theta}_f)$ around $(P^0, \alpha^0, \theta_f^0)$, applying the triangle inequality to $||\tilde{P}_{j-1} - P^0||$ and $||\tilde{\alpha}_{j-1} - \alpha^0||$, and using Lemma 7(f), $\sup_{(a,x)} \sup_{(P,\theta)} ||D^3 \ln \Psi(P, \theta)(a|x)|| < \infty$, $\sup_{(a,x)} \sup_\theta ||(\partial/\partial\theta) P_\theta(a|x)|| < \infty$, and Lemma 9.

Similarly, we obtain $\sup_{\theta^0 \in \Theta_1} \Pr_{\theta^0}(||Q_N(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})^{-1}|| > K) = o(N^{-c})$ by expanding $Q_N(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})$ around $(P^0, \alpha^0, \theta_f^0)$ and applying Lemma A.2(a) of A05 and Assumption 7(c).

In case of NR, the first term on the right of (A-5) is zero. Hence, the first equation of (A-2) for $k = j$ follows from these bounds on $D_P \overline{\psi}_\alpha(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})$ and $Q_N(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})^{-1}$. In case of the default NR, line-search NR, and OPG, repeating the argument of the proof of Lemma 1 of Andrews (2001) gives $\sup_{\theta^0 \in \Theta_1} \Pr_{\theta^0}(||D_\alpha \overline{\psi}_\alpha(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1}) - Q_N(\tilde{P}_{j-1}, \tilde{\alpha}_{j-1})|| > N^{-1/2} \ln N) = o(N^{-c})$. Using this, we can bound the first term on the right of (A-5) and establish that the first equation of (A-2) holds for $k = j$. To show that the second equation of (A-2) holds for $k = j$, expanding $\Psi(\tilde{P}_{j-1}, \tilde{\alpha}_j)$ around $(\hat{P}, \hat{\alpha})$ and applying $D_P \Psi(\hat{P}, \hat{\alpha}) = 0$ give $||\tilde{P}_j - \hat{P}|| \leq ||D_\alpha \Psi(\hat{P}, \hat{\alpha})|| ||\tilde{\alpha}_j - \hat{\alpha}|| + (\sup_{(P,\alpha)} ||D^2 \Psi(P, \alpha, \hat{\theta}_f)||) (||\tilde{\alpha}_j - \hat{\alpha}||^2 + ||\tilde{P}_{j-1} - \hat{P}||^2)$. Then the required result follows from $\sup_{(P,\theta)} ||D\Psi(P, \theta)|| < \infty$ and $\sup_{(P,\theta)} ||D^2 \Psi(P, \theta)|| < \infty$.

We proceed to prove (A-3) and (A-4). Let Σ_r denote $(\Sigma_N(\hat{\theta}))_{rr}$. Also, let $\Sigma_{k,r}$ denote Σ_r with $D_N(\hat{\theta})$ and $V_N(\hat{\theta})$ replaced with $D_N^{PL}(\tilde{P}_k, \tilde{\theta}_k)$ and $V_N^{PL}(\tilde{P}_k, \tilde{\theta}_k)$, where $\tilde{\theta}_k = (\tilde{\alpha}_k', \hat{\theta}_f')$. In view of the arguments in pp. 205-6 of A05, (A-3) holds if there exists $K < \infty$ and $\delta > 0$ such that

$$\sup_{\theta^0 \in \Theta_1} \Pr_{\theta^0}(|\Sigma_r - \Sigma_{k,r}| > \mu_{N,k}) \;=\; o(N^{-c}), \quad \text{(A-6)}$$

$$\sup_{\theta^0 \in \Theta_1} \Pr_{\theta^0}(\Sigma_{k,r} < \delta) = o(N^{-c}), \qquad \sup_{\theta^0 \in \Theta_1} \Pr_{\theta^0}(\Sigma_r < \delta) = o(N^{-c}). \quad \text{(A-7)}$$

Let $\bar{\theta}$ denote an estimator that satisfies: for all $\varepsilon > 0$, $\sup_{\theta^0 \in \Theta_1} \Pr_{\theta^0}(||\bar{\theta} - \theta^0|| > \varepsilon) = o(N^{-c})$. Then, proceeding in the same way as the proof of Lemma A.3 of A05, we obtain the following; for all $\varepsilon > 0$ and some $K < \infty, \sup_{\theta^0 \in \Theta_1} \Pr_{\theta^0}(||V_N(\bar{\theta}) - V(\theta^0)|| > \varepsilon) = o(N^{-c})$ and

$\sup_{\theta^0\in\Theta_1}\Pr_{\theta^0}(||D_N(\bar{\theta})-D(\theta^0)||>\varepsilon)=o(N^{-c})$. Thus, (A-7) holds. Equation (A-6) holds if $\sup_{\theta^0\in\Theta_1}\Pr_{\theta^0}(||V_N^{PL}(\tilde{P}_k,\tilde{\theta}_k)-V_N(\hat{\theta})||>\mu_{N,k})=o(N^{-c})$ and $\sup_{\theta^0\in\Theta_1}\Pr_{\theta^0}(||D_N^{PL}(\tilde{P}_k,\tilde{\theta}_k)-D_N(\hat{\theta})||>\mu_{N,k})=o(N^{-c})$. Note that $V_N(\hat{\theta})=V_N^{PL}(\hat{P},\hat{\theta})$ from (A-10). Therefore, the first result follows from applying the generalized Taylor's theorem to $V_N^{PL}(\tilde{P}_k,\tilde{\theta}_k)-V_N^{PL}(\hat{P},\hat{\theta})$ in conjunction with Lemma A.2(b) of A05 and (A-2). The second result is proven in an analogous manner, and we complete the proof of (A-3). Finally, in view of the argument in p. 206 of A05, (A-4) follows from (A-2) and the proof of (A-3), because Lemma A.8(a) of A05 holds in our case (see the proof of Lemma 2). The proof for the one-step NPL for general $k\geq 1$ follows by induction.

The proof for the one-step NMPL algorithm follows an analogous argument, and hence is omitted. $\square$

3 Proof of Lemmas in Appendix B

Lemma 7 collects the bounds that are used in the proof of Propositions 2-4, A.1, and Lemma 3. Lemma 8 collects the results on the derivatives of $\ln\Psi_2(P,\theta)$. Lemma 9 is our version (i.e., for $\hat{\alpha}$ and $\hat{\theta}_f$) of Lemma A.4 of A05. Lemma 10 is our version (i.e., for $\hat{\alpha}$ and $\hat{\theta}_f$) of Lemma A.6 of A05.

Lemma 7 *Suppose Assumptions 1-5 hold, $\bar{P}\rightarrow_p P^0$, and $\bar{\theta}\rightarrow_p\theta^0$. Let $\psi_i(P,\theta)$ denote either $\ln\Psi(P,\theta)(a_i|x_i)$ or $\ln\Psi_2(P,\theta)(a_i|x_i)$. Then*

$$
\begin{array}{ll}
(a) & D^s\Psi(\bar{P},\bar{\theta})(a_i|x_i)=O_p(1)\quad \text{for } s=1,2,\\
(b) & N^{-1}\sum_{i=1}^N \sup_{(P,\theta)\in B_P\times\Theta_0}||D^s\psi_i(P,\theta)||^q=O_p(1)\quad \text{for } q=1,2 \text{ and } s=1,\dots,4,\\
(c) & \sup_{h\in B_p}||N^{-1}\sum_{i=1}^N D_{P\alpha}\ln\Psi(P^0,\theta^0)(a_i|x_i)h||=O_p(N^{-1/2}),\\
(d) & N^{-1}\sum_{i=1}^N D^2\psi_i(\bar{P},\bar{\theta})=E_{\theta^0}D^2\psi_i(P^0,\theta^0)+O_p(||\bar{P}-P^0||+||\bar{\theta}-\theta^0||+N^{-1/2}),\\
(e) & \left\{\begin{array}{l} N^{-1}\sum_{i=1}^N D_\theta\psi_i(\bar{P},\bar{\theta})D_\theta\psi_i(\bar{P},\bar{\theta})\\ =E_{\theta^0}D_\theta\psi_i(P^0,\theta^0)D_\theta\psi_i(P^0,\theta^0)+O_p(||\bar{P}-P^0||+||\bar{\theta}-\theta^0||+N^{-1/2}).\end{array}\right.
\end{array}
$$

If Assumptions 1-8 hold, then (b) holds for $(P,\theta)\in B_P\times\Theta_1$.

(f) Suppose Assumptions 1-8 hold. Then, for all $\varepsilon>0$ and $c>0$,
$\sup_{\theta^0\in\Theta^1}\Pr(||N^{-1}\sum_{i=1}^N D_{P\alpha}\ln\Psi(P^0,\theta^0)(a_i|x_i)||>\varepsilon N^{-1/2}\ln N)=o(N^{-c})$.

Proof Parts (a) and (b) follow from Assumptions 4(c), 4(g), and 5(b).

For part (c), first recall $ED_{P\alpha}\ln\Psi(P^0,\theta^0)(a_i|x_i)=0$ from the information matrix equality and Proposition 1. When the support of x_i is finite, the stated result follows immediately because $D_{P\alpha}\ln\Psi(P^0,\theta^0)(a|x)$ is a matrix. When some elements of x_i are continuously distributed, we apply the framework of Section B.1 of Ichimura and Lee (2006), who build on van der Vaart and Wellner (1996) (VW hereafter). Without loss of generality, assume all the elements of x are continuously distributed. Define $y=\{a,x\}$ and $m_h(y_i)=D_{P\alpha}\ln\Psi(P^0,\theta^0)(a_i|x_i)h$. Let

$\mathcal{M} = \{m_h(y) : \ h \in B_P\}$. Then, it suffices to show $\sup_{m_h \in \mathcal{M}} |N^{-1/2} \sum_{i=1}^N m_h(y_i)| = O_p(1)$. From Theorem 2.14.2 of VW, there exists a constant C such that

$$E\left[\sup_{m_h \in \mathcal{M}} \left| N^{-1/2} \sum_{i=1}^N m_h(y_i) \right| \right] \le C \int_0^1 \sqrt{1 + \log N_{[]}(\varepsilon ||M||_{P,2}, \mathcal{M}, ||\cdot||_{P,2})} d\varepsilon ||M||_{P,2}, \quad \text{(A-8)}$$

where $N_{[]}(\varepsilon, \mathcal{M}, ||\cdot||)$ is the bracketing number for the set $\mathcal{M}$, $M(y) = \sup_{m_h \in \mathcal{M}} |m_h(y)|$, and $||M||_{P,2} = (E|M(y)|^2)^{1/2}$. See VW p. 83 for exact definitions. In our case, $||M||_{P,2} < \infty$ from Assumption 4(g). Since $m_h(y)$ is a linear operator in h, it follows from Theorem 2.7.11 of VW that $N_{[]}(2\varepsilon ||M||_{P,2}, \mathcal{M}, ||\cdot||_{P,2}) \le N(\varepsilon, B_P, ||\cdot||_\infty)$, where $N(\varepsilon, B_P, ||\cdot||_\infty)$ is the covering number for the set B_P (see VW p. 83 for the definition), and $||\cdot||_\infty$ is the sup norm in B_P. Finally, it follows from the smoothness of $P(a|x)$ specified in Assumption 4(i) and Theorem 2.7.1 of VW that $\log N(\varepsilon, B_P, ||\cdot||_\infty) \le C_K (1/\varepsilon)^\beta$ with $\beta < 2$ and $C_K < \infty$. Consequently, the left hand side of (A-8) is finite, and part (c) follows.

Parts (d) and (e) follow from part (b) and the law of large numbers.

For part (f), from Theorem 2.14.24 of VW, there exist constants C and D such that

$$\Pr\left(\sup_{m \in \mathcal{M}} \left| N^{-1/2} \sum_{i=1}^N m(y_i) \right| > Ct \right) \le D \exp - \frac{t^2 N^{1/2}}{\max(\mu_N, N^{-1/2}) + N^{1/2}\sigma_{\mathcal{M}}^2}, \quad \text{(A-9)}$$

for all t such that $\mu_N \le t \le \max(\mu_N, N^{-1/2}) + N^{1/2}\sigma_{\mathcal{M}}^2$, where $\mu_N = E[\sup_{m \in \mathcal{M}} |N^{-1/2} \sum_{i=1}^N m_h(y_i)|]$ and $\sigma_{\mathcal{M}}^2 = \sup_{m \in \mathcal{M}} |E(m - Em)^2|$. Note that $\mu_N < \infty$ from part (c) and $\sigma_{\mathcal{M}}^2 < \infty$ because m is bounded. Set $t = \varepsilon \log N / C$. Then, for sufficiently large N, $\mu_N \le t \le \max(\mu_N, N^{-1/2}) + N^{1/2}\sigma_{\mathcal{M}}^2$ holds, and the right hand side of (A-9) is bounded by $D \exp -c_2 (\log N)^2$ for a constant $c_2 > 0$, which is $o(N^{-c})$ for any $c > 0$. $\square$

Lemma 8 *Suppose Assumptions 1-4 hold. Then*

$$(a) \quad \begin{cases} D_P \ln \Psi_2(P_\theta, \theta)(a_i|x_i) = 0, \quad D_\theta \ln \Psi_2(P_\theta, \theta)(a_i|x_i) = D \ln P_\theta(a_i|x_i), \\ D_{\theta\theta} \ln \Psi_2(P_\theta, \theta)(a_i|x_i) = D^2 \ln P_\theta(a_i|x_i), \quad D_{P\theta} \ln \Psi_2(P_\theta, \theta)(a_i|x_i) = 0. \\ \textit{The same results hold for the derivatives of } \Psi_2(P_\theta, \theta)(a_i|x_i) \textit{ and } P_\theta(a_i|x_i). \end{cases}$$

$$(b) \quad E_{\theta^0} D_{PP\theta} \ln \Psi_2(P^0, \theta^0)(a_i|x_i) = 0, \quad E_{\theta^0} D_{\theta P\theta} \ln \Psi_2(P^0, \theta^0)(a_i|x_i) = 0.$$

$$(c) \quad \begin{cases} \sup_{(h_1,h_2) \in B_P \times B_P} ||N^{-1} \sum_{i=1}^N D_{PP\theta} \ln \Psi_2(P^0, \theta^0)(a_i|x_i) h_1 h_2|| = O_p(N^{-1/2}), \\ \sup_{(h_1,h_2) \in \Theta \times B_P} ||N^{-1} \sum_{i=1}^N D_{\theta P\theta} \ln \Psi_2(P^0, \theta^0)(a_i|x_i) h_1 h_2|| = O_p(N^{-1/2}). \end{cases}$$

Proof The first result of part (a) is a simple consequence of Proposition 1 and the chain rule. For the other results of part (a), recall $P_\theta(a_i|x_i)$ is defined implicitly as a function of θ as $P_\theta(a_i|x_i) = \Psi(P_\theta, \theta)(a_i|x_i)$. Taking the derivative of $\ln P_\theta(a_i|x_i) = \ln \Psi(P_\theta, \theta)(a_i|x_i)$ and using Proposition 1 gives

$$D \ln P_\theta(a_i|x_i) = D_P \ln \Psi(P_\theta, \theta)(a_i|x_i) D P_\theta + D_\theta \ln \Psi(P_\theta, \theta)(a_i|x_i) = D_\theta \ln \Psi(P_\theta, \theta)(a_i|x_i). \quad \text{(A-10)}$$

It follows from the chain rule and $D_P\Psi(P_\theta,\theta)=0$ that, for all $h\in\Theta$,

$$\begin{aligned} D^2\ln P_\theta(a_i|x_i)h &= D_{PP}\ln\Psi(P_\theta,\theta)(a_i|x_i)DP_\theta h\cdot DP_\theta + D_{\theta P}\ln\Psi(P_\theta,\theta)(a_i|x_i)h\cdot DP_\theta \\ &\quad + D_{P\theta}\ln\Psi(P_\theta,\theta)(a_i|x_i)\cdot DP_\theta h + D_{\theta\theta}\ln\Psi(P_\theta,\theta)(a_i|x_i)h. \end{aligned} \tag{A-11}$$

Now collect the derivatives of $\ln\Psi_2(P,\theta)=\ln\Psi(\Psi(P,\theta),\theta)$, where P is not necessarily the fixed point of $\Psi(\cdot,\theta)$.

$$D_\theta\ln\Psi_2(P,\theta)(a_i|x_i) = D_P\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)D_\theta\Psi(P,\theta)+D_\theta\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i), \tag{A-12}$$

where $D_P\ln\Psi(\Psi(P,\theta),\theta)$ is the F-derivative of $\ln\Psi(P,\theta)$ with respect to P evaluated at $(\Psi(P,\theta),\theta)$, and similarly for $D_{PP}\ln\Psi(\Psi(P,\theta),\theta)$ etc. Furthermore, for all $h\in\Theta$

$$\begin{aligned} D_{\theta\theta}\ln\Psi_2(P,\theta)(a_i|x_i)h &= D_{PP}\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)D_\theta\Psi(P,\theta)h\cdot D_\theta\Psi(P,\theta) \\ &\quad + D_{\theta P}\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)h\cdot D_\theta\Psi(P,\theta) + D_P\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)D_{\theta\theta}\Psi(P,\theta)h \\ &\quad + D_{P\theta}\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)D_\theta\Psi(P,\theta)h + D_{\theta\theta}\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)h. \end{aligned} \tag{A-13}$$

The cross derivative of $\Psi_2(P,\theta)$ takes the form, for all $h\in B_P$

$$\begin{aligned} D_{P\theta}\ln\Psi_2(P,\theta)(a_i|x_i)h &= D_{PP}\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)D_P\Psi(P,\theta)h\cdot D_\theta\Psi(P,\theta) \\ &\quad + D_P\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)D_{P\theta}\Psi(P,\theta)h + D_{P\theta}\ln\Psi(\Psi(P,\theta),\theta)(a_i|x_i)D_P\Psi(P,\theta)h. \end{aligned} \tag{A-14}$$

Evaluating (A-12)-(A-14) at $P=P_\theta$ with $D_P\Psi(P_\theta,\theta)=0$ and using (A-10)-(A-11) gives the first set of the results in part (a). The required results for the derivatives of $\Psi_2(P_\theta,\theta)(a_i|x_i)$ and $P_\theta(a_i|x_i)$ follow from the same argument.

To show part (b), taking the F-derivative of (A-14) and evaluating it at $P=P_\theta$ gives, for all $h_1,h_2\in B_P$, $D_{PP\theta}\ln\Psi_2(P_\theta,\theta)(a_i|x_i)h_1h_2 = D_{PP}\ln\Psi(P_\theta,\theta)(a_i|x_i)D_{PP}\Psi(P_\theta,\theta)h_1h_2\cdot D_\theta\Psi(P_\theta,\theta)+D_{P\theta}\ln\Psi(P_\theta,\theta)(a_i|x_i)D_{PP}\Psi(P_\theta,\theta)h_1h_2$. Similarly, for all $k_1\in\Theta$ and $k_2\in B_P$, $D_{\theta P\theta}\ln\Psi_2(P_\theta,\theta)(a_i|x_i)k_1k_2 = D_{PP}\ln\Psi(P_\theta,\theta)(a_i|x_i)\ D_{\theta P}\Psi(P_\theta,\theta)k_1k_2\cdot D_\theta\Psi(P_\theta,\theta)$ $+D_{P\theta}\ln\Psi(P_\theta,\theta)(a_i|x_i)D_{\theta P}\Psi(P_\theta,\theta)k_1k_2$. Part (b) follows because $E_{\theta^0}D_{PP}\ln\Psi(P^0,\theta^0)(a_i|x_i)=0$ and $E_{\theta^0}D_{P\theta}\ln\Psi(P^0,\theta^0)(a_i|x_i)=0$ from Proposition 1 and the information matrix equality.

The proof of part (c) follows from the same argument as the proof of part (c) of Lemma 7. The only difference is that $D_{PP\theta}$ is an operator in $h,k\in B_P\times B_P$, which has $2d$ continuously distributed elements. $\square$

Lemma 9 *Suppose Assumptions 1-8 hold. Then, for all* $\varepsilon>0$,

$$\sup_{\theta^0\in\Theta^1}\Pr_{\theta^0}\left(N^{1/2}||\hat\theta_f-\theta^0_f||+N^{1/2}||\hat\alpha-\alpha^0||>\varepsilon\ln N\right)=o(N^{-c}).$$

Proof From Lemma 5 of Andrews (2001), we have $\sup_{\theta^0_f\in\Theta^1_f}\Pr_{\theta^0_f}(N^{1/2}||\hat\theta_f-\theta^0_f||>\varepsilon\ln N)=o(N^{-c})$ for all $\varepsilon>0$.

Define $\rho_N(\alpha,\theta_f) = -N^{-1}\sum_{i=1}^N \ln P_{(\alpha,\theta_f)}(a_i|x_i)$ and $\rho(\alpha,\theta_f) = -E_{\theta^0}\ln P_{(\alpha,\theta_f)}(a_i|x_i)$, so that $\hat{\alpha} = \arg\min_{\alpha\in\Theta_\alpha}\rho_N(\alpha,\hat{\theta}_f)$. By Assumption 6(b), given any $\epsilon > 0$, there exists $\delta > 0$ such that $||\alpha - \alpha^0|| > \varepsilon$ implies $\rho(\alpha,\theta_f^0) - \rho(\alpha^0,\theta_f^0) \geq \delta$. Therefore, $\sup_{\theta^0\in\Theta^1}\Pr_{\theta^0}(||\hat{\alpha} - \alpha^0|| > \varepsilon) \leq \sup_{\theta^0\in\Theta^1}\Pr_{\theta^0}(\rho(\hat{\alpha},\theta_f^0) - \rho(\alpha^0,\theta_f^0) \geq \delta)$. Since $\rho(\alpha,\theta_f)$ is uniformly continuous, the right hand is no larger than $\sup_{\theta^0\in\Theta^1}\Pr_{\theta^0}(\rho(\hat{\alpha},\hat{\theta}_f) - \rho(\alpha^0,\hat{\theta}_f) \geq \delta/2) + o(N^{-c}) \leq \sup_{\theta^0\in\Theta^1}\Pr_{\theta^0}(\rho(\hat{\alpha},\hat{\theta}_f) - \rho_N(\hat{\alpha},\hat{\theta}_f) + \rho_N(\alpha^0,\hat{\theta}_f) - \rho(\alpha^0,\hat{\theta}_f) \geq \delta/2) + o(N^{-c}) = o(N^{-c})$, where the first inequality follows from $\rho_N(\hat{\alpha},\hat{\theta}_f) - \rho_N(\alpha^0,\hat{\theta}_f) \leq 0$ and the last equality follows from $\sup_{\theta^0\in\Theta^1}\Pr_{\theta^0}(\sup_{(\alpha,\theta_f)\in\Theta}|\rho_N(\alpha,\theta_f) - \rho(\alpha,\theta_f)| > \eta) = o(N^{-c})$ for all $\eta > 0$, which follows from (8.49) in Andrews (2001).

Therefore, we can use the argument in p. 34 of Andrews (2001) following his equation (8.51) to obtain $\inf_{\theta^0\in\Theta^1}\Pr_{\theta^0}((\partial/\partial\alpha)\rho_N(\hat{\alpha},\hat{\theta}_f) = 0) = 1 - o(N^{-c})$. Then, the stated result for $\hat{\alpha}$ follows from expanding $(\partial/\partial\alpha)\rho_N(\hat{\alpha},\hat{\theta}_f)$ around (α^0,θ_f^0) and applying an argument similar to (8.52) in Andrews (2001). $\square$

Lemma 10 *Suppose Assumptions 1-8 hold. Define $S_N(\theta) = N^{-1}\sum_{i=1}^N h(w_i,\theta)$ and $\hat{\theta} = (\hat{\alpha}',\hat{\theta}_f')'$. Let $\Delta_N(\theta^0)$ denote $N^{1/2}(\hat{\theta} - \theta^0)$, $T_N(\theta_r^0)$, or $H_N(\hat{\theta},\theta^0)$. Let L denote the dimension of $\Delta_N(\theta^0)$. For each definition of $\Delta_N(\theta^0)$, there is an infinitely differentiable function $G(\cdot)$ that does not depend on θ^0 and that satisfies $G(E_{\theta^0}S_N(\theta^0)) = 0$ for all N large and all $\theta^0 \in \Theta_1$, and $\sup_{\theta^0\in\Theta_1}\sup_{B\in\mathcal{B}_L}\left|\Pr_{\theta^0}(\Delta_N(\theta^0)\in B) - \Pr_{\theta^0}(N^{1/2}G(S_N(\theta^0))\in B)\right| = o(N^{-c})$, where $\mathcal{B}_L$ denotes the class of all convex sets in $\mathbb{R}^L$.*

Proof The proof follows the proof of Lemma A.6 of A05. Suppose $\Delta_N(\theta^0) = N^{1/2}(\hat{\theta} - \theta^0)$. Define $s(\theta) = [(\partial/\partial\alpha')N^{-1}\sum_{i=1}^N \ln P_{(\alpha,\theta_f)}(a_i|x_i), (\partial/\partial\theta_f')N^{-1}\sum_{i=1}^N \ln f_{\theta_f}(x_i'|a_i,x_i)]'$. From Lemma 9, $\hat{\theta}$ is in the interior of Θ with probability $1 - o(N^{-c})$, and we have $\inf_{\theta^0\in\Theta_1}\Pr_{\theta^0}(s(\hat{\theta}) = 0) = 1 - o(N^{-c})$. Consequently, the proof of Lemma A.6 of A05 carries through if we replace $(\partial/\partial\theta)\rho_N(\theta)$ and $\hat{\theta}_N$ in A05 with our $s(\theta)$ and $\hat{\theta}$. The only difference is $(\partial/\partial x)\nu(E_{\theta_0}R_N(\theta_0),x)|_{x=0} = N^{-1}\sum_{i=1}^N E_{\theta_0}g(\tilde{W}_i,\theta_0)g(\tilde{W}_i,\theta_0)'$ in line 20, p. 210 of A05 needs to be replaced with

$$\frac{\partial}{\partial x}\nu(E_{\theta^0}R_N(\theta^0),x)|_{x=0} = E\begin{bmatrix} (\partial^2/\partial\alpha\partial\alpha')\ln P_{\theta^0}(a_i|x_i) & (\partial^2/\partial\alpha\partial\theta_f')\ln P_{\theta^0}(a_i|x_i) \\ 0 & (\partial^2/\partial\theta_f\partial\theta_f')\ln f_{\theta_f^0}(x_i'|a_i,x_i) \end{bmatrix}.$$

Because this is negative definite, the implicit function theorem can be applied to $\nu(\cdot,\cdot)$ at the point $(E_{\theta^0}R_N(\theta^0),0)$, to obtain $\inf_{\theta^0\in\Theta_1}\Pr_{\theta^0}(\hat{\theta} - \theta^0 = \Lambda(R_N(\theta^0) + e_N(\theta^0))) = 1 - o(N^{-c})$. This equation corresponds to (A.35) of A05, where $R_N(\theta^0)$ and $e_N(\theta^0)$ are defined in the same manner as in A05 but his $(\partial/\partial\theta)\rho_N(\theta_0)$ replaced with our $s(\theta^0)$. The remaining part of his proof carries through, because Lemmas A.5 and A.8 of A05 holds in our context by our Assumptions 1-8, and our Lemma 9 plays the role of Lemma A.4 of A05. $\square$

References

[1] Andrews, D. W. K. (2001). Higher-order improvements of the parametric bootstrap for Markov processes. Cowles Foundation Discussion Paper No. 1334, Yale University.

[2] Andrews, D. W. K. (2005). "Higher-order improvements of the parametric bootstrap for Markov processes." in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, eds. D. W. K. Andrews and J. H. Stock, Cambridge University Press.

[3] Ichimura, H. and S. Lee (2006). Characterization of the asymptotic distribution of semiparametric M-estimators. Mimeographed, University of Tokyo.

[4] van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes.* New York, Springer-Verlag.