Transition Probabilities and Moment Restrictions in Dynamic Fixed Effects Logit Models *

Kevin Dano
Princeton University

November 5, 2025

Abstract

This paper introduces a new method to derive moment restrictions in dynamic discrete choice models with strictly exogenous regressors, fixed effects and logistic errors. We show how the structure of logit probabilities and basic properties of rational fractions can be used to construct moment functions free of the fixed effects in a way that scales automatically with the lag order and the number of observed periods. We demonstrate the approach in binary response models of arbitrary lag order, first-order panel vector autoregressions and dynamic multinomial logit models. The semiparametric efficiency bound is characterized for the leading binary case with one lag. Finally, we illustrate our results in an application investigating the dynamics of drug consumption among young people.

Keywords: dynamic discrete choice, panel data, fixed effects.
JEL Classification Codes: C23, C33.

*E-mail: `kdano@princeton.edu`. I am very grateful to my advisors Bryan Graham, Stéphane Bonhomme, Demian Pouzo and Jim Powell for their generous support and advice. I also thank Chris Muris, Yassine Sbai Sassi, Bo Honoré, Martin Weidner, Adam Rosen, Stefan Hoderlein, Elena Manresa, Michal Kolesár, Ulrich Müller, Bocar Ba, Nick Gebbia, Tahsin Saffat, seminar participants at UCSD, Emory, NYU, Zurich, LSE, Michigan, Duke, UPenn, Princeton, UCLA, Harvard, MIT, Penn State, Georgetown, Warwick, UCL and the audiences at the 2025 SITE conference, the 2023 IAAE Annual Conference, the 2023 California Econometrics Conference and the 2023 Causal Panel Data Conference at Stanford GSB for valuable comments and discussions. Financial support from the 2023 IAAE Conference is gratefully acknowledged. All errors are my own.

1 Introduction

Dynamic discrete choice models with logistic errors and unobserved individual heterogeneity underlie much work examining state-dependence in economics. Examples include studies of labor market outcomes (Magnac (2000)), welfare participation (Chay et al. (1999), Card and Hyslop (2005)), health plan choices (Pakes et al. (2021)), drug addiction (Deza (2015)), and even transitivity in networks (Graham (2013), Graham (2016)). By nature, inference in such models can be complex, but a powerful principle is to look for orthogonality restrictions independent of unit-specific effects to secure consistent estimation of common parameters. In short panels, these so-called fixed effects strategies bypass two issues: i) the *incidental parameters problem* associated with maximum likelihood estimation (Neyman and Scott (1948)), ii) risking misspecification by parameterizing individual heterogeneity and its relationship with initial outcomes which are inherently unknown.

An early breakthrough providing restrictions of this type in simple models with only a lagged outcome variable came from conditional likelihood, as exemplified by Cox (1958), Chamberlain (1985), Magnac (2000). This approach leverages sufficient statistics tied to the logistic assumption to eliminate the fixed effect. Subsequently, Honoré and Kyriazidou (2000) extended this idea to models with strictly exogenous regressors, showing its viability if the regressors remain constant over specific time periods (see also, Honoré and Kyriazidou (2019), Muris et al. (2020)). While relevant in certain settings, the stability requirement on the regressors means two limitations for the conditional likelihood approach: it inherently rules out time effects and implies rates of convergence slower than $\sqrt{N}$ for continuous explanatory variables. Moreover, calculations from Honoré and Kyriazidou (2000) suggested that it does not easily extend to models with a higher lag order. These shortcomings have motivated the search for alternative solutions, culimating in a recent moment-based paradigm. Its essence is the construction of moment functions free from fixed effects, beyond scores of conditional likelihoods, and enabling general estimation at $\sqrt{N}$-rate. Kitazawa et al. (2013, 2016) and Kitazawa (2022) represent creative examples of this idea for the AR(1) - autoregressive of order one - logit model. A more systematic framework to obtain moment restrictions is offered by functional differencing (Bonhomme (2012)), and the recent contributions of Honoré and Weidner (2020), Honoré et al. (2021), and to some extent Dobronyi et al. (2021)[1] can be viewed as powerful displays of

[1]Dobronyi et al. (2021) also derive moment inequality conditions in AR(1) and AR(2) logit models which is beyond the scope of standard functional differencing.

this technique in discrete choice models when coupled with symbolic computing (e.g Mathematica).

The core contribution of this paper is a new general approach to construct moment restrictions in a broad class of dynamic fixed effects logit models (henceforth DFEL), where unit-specific effects feature as "heterogeneous" intercepts. This class encompasses many specifications commonly encountered in applications but excludes models with heterogeneous coefficients on lagged outcomes and/or regressors as in Chamberlain (1985) and Browning and Carro (2014). Unlike recent competing methods, ours does not require numerical experimentation or symbolic computing, enabling us to advance on theoretical and operational fronts. First, we show that the existence of moment restrictions in DFEL models is rooted in the rational fractional structure of logit probabilities with respect to fixed effects. Fundamentally, this is because products of rational fractions can be decomposed into sums of simpler rational fractions. Leveraging this fact, we formally resolve open conjectures regarding the number of moment conditions available in binary logit models. Second, our procedure scales efficiently with the number of time periods, and also with the lag order in binary response models. A key result is the discovery of a novel recursive formula that enables the construction of moment restrictions for an AR(p) from features of an AR($p-1$). Third, the algebraic foundation of our procedure allows us to easily derive extensions for the VAR(1) logit model, the dynamic multinomial logit model, and dynamic network formation models in the spirit of Graham (2013), Graham (2016), where symbolic computation quickly encounters tractability limits. Detailed results for the latter two models are provided in the Supplemental Appendix.

The method exploits two key insights. First, the (individual-specific) transition probabilities of logit models can often be expressed as conditional expectations of functions of observables and common parameters given the initial condition, the regressors and the fixed effects. We refer to these moment functions as *transition functions*. They have the crucial feature of not depending on individual fixed effects. Second, with sufficient time periods, many transition probabilities admit at least two distinct transition functions. Together, these insights motivate a natural two-step recipe to systematically form valid moment functions: **Step 1)** compute the model's transition functions, **Step 2)** take differences of two transition functions associated to the same transition probability. We find that a careful application of this procedure yields all the moment equality restrictions available for binary response models. We build on this property to characterize the efficiency bound in the leading AR(1) logit model, complementing Hahn (2001) and Gu et al. (2023).

The remainder of the paper proceeds as follows. In Section 2, we introduce the class of models under consideration and outline our methodology for obtaining moment restrictions. Section 3 gives a detailed analysis of the baseline AR(1) logit model. We present a new perspective to enumerate the available moment restrictions, demonstrate our approach for deriving their expressions, and characterize the efficiency bound. In Section 4 and Section 5, we provide some extensions for AR(p) logit models and the VAR(1) logit model. Section 6 contains an empirical application investigating the dynamics of drug consumption among young people. Section 7 concludes. The Appendix contains proofs of key results. The Supplemental Appendix compiles auxiliary results and discussions of the dynamic multinomial logit model and a dynamic network formation model, which may be of independent interest.

2 Setup, objective, and methodology

General setup. The setting is panel data with $i = 1, \ldots, N$ individuals followed over $t = 1, \ldots, T$ periods. The econometrician observes (Y_i^0, Y_i, X_i) for all individuals, where $Y_i = (Y_{i1}, \ldots, Y_{iT}) \in \mathcal{Y}^T$ denotes the endogenous discrete outcomes, $X_i = (X_{i1}, \ldots, X_{iT}) \in \mathcal{X}^T$ the covariates and $Y_i^0 = (Y_{i0}, Y_{i-1}, \ldots)$ the initial condition, i.e the set of observed outcomes prior to period $t = 1$. The models we are considering feature two components. The first component is a parametric model of outcomes Y_i conditional on strictly exogenous explanatory variables (Y_i^0, X_i) and time-invariant unobserved heterogeneity $A_i \in \mathcal{A}$. For a known lag order $p \geq 1$, and using the shorthand $z_s^t = (z_t, z_{t-1}, \ldots, z_s)$ for $s < t$, it takes the form

$$f(y|y^0, x, a) = P(Y_i = y|Y_i^0 = y^0, X_i = x, A_i = a) = \prod_{t=0}^{T-1} \pi_t^{y_{t+1}|y_{t-(p-1)}^t}(a, x; \theta_0)$$

where $\pi_t^{y_{t+1}|y_{t-(p-1)}^t}(a, x; \theta_0) = P(Y_{it+1} = y_{t+1}|Y_{it-(p-1)}^t = y_{t-(p-1)}^t, X_i = x, A_i = a)$ denote the model's (individual-specific) transition probabilities known up to the finite dimensional parameter θ_0. We shall omit the dependence on θ_0 in the sequel. The second component of the model is the distribution of heterogeneity A_i conditional on $(Y_i^0 = y^0, X_i = x)$ which we denote as $q(.|y^0, x)$. Following a large literature in panel data, we leave it unrestricted thereby treating A_i as a "fixed effect". Jointly, the two model components map to conditional outcome probabilities

$$f(y|y^0, x) = P(Y_i = y|Y_i^0 = y^0, X_i = x) = \int_{\mathcal{A}} f(y|y^0, x, a)q(a|y^0, x)da$$

that are identified in the population. It is assumed that (Y_i^0, Y_i, X_i, A_i) is i.i.d across individuals.

Objective. We are primarily concerned with the identification and estimation of θ_0 in short panels, i.e for fixed T. To this end, the chief objective of this paper is to show how to construct moment functions $\psi_\theta(Y_i, Y_i^0, X_i)$ free of the fixed effect parameter that are valid in the sense that:

$$\mathbb{E}\left[\psi_{\theta_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i, A_i\right] = 0 \tag{1}$$

When this is possible, the law of iterated expectations implies the conditional moment:

$$\mathbb{E}\left[\psi_{\theta_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i\right] = 0$$

which can in turn be leveraged to assess the identifiability of θ_0 and form the basis of an estimation strategy by GMM or empirical likelihood[2]. This is the central idea underlying functional differencing (Bonhomme (2012)) and was recently applied by Honoré and Weidner (2020) to derive valid moment conditions for a class of dynamic logit models with scalar fixed effects. We borrow the same insight but instead of searching for solutions numerically on a case-by-case basis as explained in Honoré and Weidner (2020), we propose a complementary systematic algebraic procedure to recover the model's valid moments that we outline in the next paragraph[3].

Methodology. We call a *transition function* associated to a transition probability $\pi_t^{y_{t+1}|y_{t-(p-1)}^t}(A_i, X_i)$ any moment function $\phi_\theta(Y_i, Y_i^0, X_i)$ satisfying:

$$\mathbb{E}\left[\phi_{\theta_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i, A_i\right] = \pi_t^{y_{t+1}|y_{t-(p-1)}^t}(A_i, X_i) \tag{2}$$

In panels of sufficient length, transition functions happen to exist for certain transition probabilities in several DFEL models of interest and are typically non-unique. This non-uniqueness motivates a two-step approach to obtain valid moment functions fulfilling (1). In **Step 1)**, the researcher computes the model's transition functions. Foreshadowing results for the binary case with p lags (e.g AR(p) logit models), a minimum of $T = p + 1$ periods will generally be required to obtain unique transition

[2]Notice that for $\mathbb{E}\left[\psi_{\theta_0}(Y_i, Y_i^0, X_i) \,|\, Y_i^0, X_i\right] = 0$ to hold irrespective of the distribution of the fixed effect, (1) must be satisfied. If (1) were strictly positive on a set of positive Lebesgue measure, there would exist distributions of fixed effects q supported on that set inducing violations of the desired moment equality. The same holds true if (1) were instead strictly negative on a set of positive Lebesgue measure.

[3]We refer readers to Dobronyi et al. (2021) and Kitazawa (2022) for alternative algebraic approaches. The first paper uses the full likelihood and focuses on the AR(1) and instances of the AR(2) model. The second paper has a transformation approach adapted to the AR(1) model.

functions for a subset of transition probabilities in period $t = p$. However, this alone does not yield moment equality restrictions on θ_0, for which an additional period is necessary. With $T \geq p+2$, we explain how to systematically construct distinct transition functions associated to the same subset of transition probabilities across periods $t \in \{p+1, \ldots, T-1\}$. The key ingredient is the use of *partial fraction decompositions* for rational fractions tailored to the structure of logit transition probabilities (see Appendix Lemmas 6-7). This leads us to **Step 2)** where we simply take differences of two transition functions associated to the same transition probability to automatically obtain valid moment functions.

Intuitively, this two-step strategy emulates familiar fixed effects differencing schemes in panel data models with strict exogeneity. That is finding two moment functions whose conditional expectations given (Y_i^0, X_i, A_i) produce the same function of the fixed effects $h(A_i, X_i)$ and taking their difference. The relevant choices of $h(A_i, X_i)$ are inherently model specific but in binary logit models, any such function happens to be a linear combination of transition probabilities (see Theorems 1 and 3). This insight explains our particular focus on transition functions and transition probabilities, albeit alternative choices of "basis" are possible.

Notations. We reemphasize the use of the shorthand $Z_{is}^t = (Z_{it}, Z_{it-1}, \ldots, Z_{is})$ to denote the history of Z_{it} between periods s and t. We let Δ denote the first-differencing operator so that $\Delta Z_{it} = Z_{it} - Z_{it-1}$ and make use of the notation $Z_{its} = Z_{it} - Z_{is}$ for $s \neq t$ to accommodate long differences. We use $\mathbb{1}\{.\}$ for the indicator function; $\mathrm{Im}(f)$, $\ker(f)$, $\mathrm{rank}(f)$ to denote the image, the nullspace and the rank of a linear map f.

3 The AR(1) logit model

We begin our analysis with the textbook AR(1) logit model with fixed effects

$$Y_{it} = \mathbb{1}\{\gamma_0 Y_{it-1} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\}, \quad t = 1, \ldots, T \tag{AR1}$$

Here, $\mathcal{Y} = \{0,1\}$, $\mathcal{X} \subseteq \mathbb{R}^{K_x}$, $\mathcal{A} = \mathbb{R}$, $\theta_0 = (\gamma_0, \beta_0') \in \mathbb{R} \times \mathbb{R}^{K_x}$, and $Y_i^0 = Y_{i0}$. The logistic assumption on ϵ_{it} implies the transition probabilities

$$\pi_t^{0|0}(A_i, X_i) = P(Y_{it+1} = 0 | Y_{it} = 0, X_i, A_i) = \frac{1}{1 + e^{A_i + X_{it+1}'\beta_0}}$$
$$\pi_t^{1|1}(A_i, X_i) = P(Y_{it+1} = 1 | Y_{it} = 1, X_i, A_i) = \frac{e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}}{1 + e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}}$$

with $\pi_t^{1|0}(A_i, X_i), \pi_t^{0|1}(A_i, X_i)$ redundant since $\pi_t^{k|l}(A_i, X_i) = 1 - \pi_t^{l|l}(A_i, X_i)$ for all $(k, l) \in \mathcal{Y}^2$.

3.1 The number of moment restrictions in the AR(1)

We start out by enumerating the moment functions implied by the model that are linearly independent. This will provide a means to assess the exhaustiveness of our two-step approach. To this end, let $\mathcal{E}_{y_0,x,T}$ denote the conditional expectation operator mapping any function of the outcome variable Y_i to its conditional expectation given $Y_{i0} = y_0, X_i = x$ and the fixed effect A_i, i.e

$$\begin{aligned} \mathcal{E}_{y_0,x,T} \colon \mathbb{R}^{\mathcal{Y}^T} &\longrightarrow \mathbb{R}^{\mathbb{R}} \\ \phi(., y_0, x) &\longmapsto \mathbb{E}\left[\phi(Y_i, y_0, x) | Y_{i0} = y_0, X_i = x, A_i = .\right] \end{aligned}$$

$\mathcal{E}_{y_0,x,T}$ is one formulation of the parametric component of the model in that for any $y \in \mathcal{Y}^T$, $\mathcal{E}_{y_0,x,T}\left[\mathbb{1}\{. = y\}\right]$ yields the conditional probability of observing history y for all possible values of the fixed effect, i.e: $\mathcal{E}_{y_0,x,T}\left[\mathbb{1}\{. = y\}\right] = P(Y_i = y | Y_{i0} = y_0, X_i = x, A_i = .)$ where $P(Y_i = y | Y_{i0} = y_0, X_i = x, A_i = a) = \prod_{t=1}^{T} \frac{e^{y_t(\gamma_0 y_{t-1} + x_t'\beta_0 + a)}}{1 + e^{\gamma_0 y_{t-1} + x_t'\beta_0 + a}}, \quad \forall a \in \mathbb{R}$. We have the following result,

Theorem 1. *Consider model (AR1) with $T \geq 1$, initial condition $y_0 \in \mathcal{Y}$ and covariates $x \in \mathcal{X}^T$. Suppose that for any $t, s \in \{1, \ldots, T-1\}$ and $(y, \tilde{y}) \in \mathcal{Y}^2$, $\gamma_0 y + x_t'\beta_0 \neq \gamma_0 \tilde{y} + x_s'\beta_0$ if $t \neq s$ or $y \neq \tilde{y}$. Then, the family $\mathcal{F}_{y_0,x,T} = \left\{1, \pi_0^{y_0|y_0}(., x), (\pi_t^{0|0}(., x), \pi_t^{1|1}(., x))_{t=1}^{T-1}\right\}$ of size $2T$ forms a basis of $\mathrm{Im}(\mathcal{E}_{y_0,x,T})$ and $\dim\left(\ker(\mathcal{E}_{y_0,x,T})\right) = 2^T - 2T$.*

Theorem 1 establishes that the linear span of transition probabilities provides a minimal description of the parametric part of the model: 2^T histories are possible but their conditional probabilities can all be written with just $2T$ basis elements. This follows from the observation that when the quantity $\gamma_0 y_{t-1} + x_t'\beta_0$ in each transition probability differs across time periods [4], the conditional probability of each history $y \in \mathcal{Y}^T$ is a ratio of polynomials in $\exp(a)$, where the numerator has lower degree than the denominator, and the later is a product of distinct irreducible terms. A sufficient condition for this is that $\gamma_0 \neq 0$ and that one regressor is continuously distributed with non-zero slope. In turn, standard results on partial fraction decompositions ensure that this ratio can be expressed as a unique linear combination of transition probabilities. This implies $\mathrm{Im}(\mathcal{E}_{y_0,x,T}) \subseteq \mathcal{F}_{y_0,x,T}$. To establish the reverse inclusion, we leverage upcoming results that prove that the transition probabilities live in $\mathrm{Im}(\mathcal{E}_{y_0,x,T})$ as expectations of transition functions.

Importantly, since $\ker(\mathcal{E}_{y_0,x,T})$ is the set of valid moment functions verifying equation (1), Theorem 1 tells us that the AR(1) model features $2^T - 2T$ linearly independent moment restrictions in general. This is a consequence of the *rank nullity theorem*. The fact that $2^T - 2T$ moment conditions are available for the AR(1) appeared initially as a conjecture in Honoré and Weidner (2020) and was later established by Kruiniger (2020) and Dobronyi et al. (2021) using different arguments from here. They did not emphasize the role of the transition probabilities. Our ideas extend naturally to the case of arbitrary lags - since the transition probabilities remain rational fractions - which was hitherto unresolved. We discuss this extension in Subsection 4.1.

Remark 1 (Counting moments in logit models)**.** Decomposing conditional probabilities of choice histories into a basis can be a useful device to infer a lower bound on the number of moment restrictions in logit models. Furthermore, if these basis elements are shown to belong to the image of the conditional expectation operator, this lower bound equals the exact number of moment restrictions.

- In the static panel logit model of Rasch (1960), $\gamma_0 = 0$ and we have $\pi_t^{1|1}(.,x) = 1 - \pi_t^{0|0}(.,x)$. Thus, provided that $x_t'\beta_0 \neq x_s'\beta_0$ for all $t \neq s$, $\mathcal{F}_{x,T} = \left\{1, (\pi_t^{0|0}(.,x))_{t=0}^{T-1}\right\}$ spans the image of the conditional expectation operator. This implies at least $2^T - (T+1)$ moment restrictions. In fact, this is

[4] This condition may be violated if for example $\gamma_0 \neq 0$ but $x_t'\beta_0 = x_s'\beta_0$. However, if we let $\mathcal{I}_t = \{s \neq t : x_t'\beta_0 = x_s'\beta_0\}$, one can show using similar arguments on rational fractions that $\left\{\pi_s^{0|0}(a,x), \pi_s^{1|1}(a,x)\right\}_{s\in\mathcal{I}_t}$ will be replaced by $\left\{\pi_t^{0|0}(a,x)^j, \pi_t^{0|1}(a,x)^j\right\}_{j=2}^{|\mathcal{I}_t|}$ in the family $\mathcal{F}_{y_0,x,T}$ of Theorem 1. Since $|\mathcal{F}_{y_0,x,T}|$ is unchanged, the number of linearly independent moment functions is unchanged.

precisely the total number of moment restrictions by Remark 2 which gives the transition functions associated to each element of $\mathcal{F}_{x,T}$.

- In the Cox (1958) model, $\gamma_0 \neq 0$ but $\beta_0 = 0$ and the transition probabilities are $\pi^{0|0}(a) = \frac{1}{1+e^a}$ and $\pi^{1|1}(a) = \frac{e^{\gamma_0+a}}{1+e^{\gamma_0+a}}$ (or equivalently $\pi^{0|1}(a) = \frac{1}{1+e^{\gamma_0+a}}$). In this case, the family $\mathcal{F}_{y_0,T} = \left\{1, \left(\pi^{0|0}(.)^j, \pi^{0|1}(.)^j\right)_{j=1}^{T-1}, \pi^{0|y_0}(.)^T\right\}$ which consists of powers of the time-invariant transition probabilities spans the image of the conditional expectation operator. Since $|\mathcal{F}_{y_0,T}| = 2T$, the model produces at least $2^T - 2T$ linearly independent moment restrictions.

Having clarified the total count of moment restrictions in the AR(1) logit model, we next discuss how to construct them with our two-step procedure.

3.2 Construction of moment restrictions in the AR(1)

3.2.1 Intuition from the case with no regressors

We first explain our approach for the simple pure AR(1) model

$$Y_{it} = \mathbb{1}\{\gamma_0 Y_{it-1} + A_i - \epsilon_{it} \geq 0\}, \quad t = 1, \dots, T \tag{AR1 pure}$$

studied by Cox (1958), Chamberlain (1985) and Magnac (2000). These papers established the identification of γ_0 for $T \geq 3$ via conditional likelihood based on the insight that $(Y_{i0}, \sum_{t=1}^{T-1} Y_{it}, Y_{iT})$ are sufficient statistics for the fixed effect. Our methodology is conceptually different as we seek to directly construct moment functions verifying equation (1).

Here, the transition probabilities are time invariant and given by

$$\pi^{k|l}(A_i) = P(Y_{it+1} = k | Y_{it} = l, A_i) = \frac{e^{k(\gamma_0 l + A_i)}}{1 + e^{\gamma_0 l + A_i}}, \quad \forall (l, k) \in \mathcal{Y}$$

Step 1). We begin by deriving the transition functions for $\pi^{0|0}(A_i)$ and $\pi^{1|1}(A_i)$. A natural starting place is to investigate the case $T = 2$, i.e 2 periods of observations after the initial condition. Recalling definition (2), we search for $\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0})$, respectively $\phi_\theta^{1|1}(Y_{i2}, Y_{i1}, Y_{i0})$, whose conditional expectation given (Y_{i0}, A_i) yields $\pi^{0|0}(A_i)$, respectively $\pi^{1|1}(A_i)$. For the purposes of illustration, let us derive

$\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0})$ step by step. By Bayes's rule:

$$\begin{aligned}
&\mathbb{E}\left[\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0}) \mid Y_{i0} = y_0, A_i = a\right] \\
&= \sum_{y_2=0}^{1}\sum_{y_1=0}^{1} P(Y_{i2} = y_2|Y_{i1} = y_1, A_i = a)P(Y_{i1} = y_1|Y_{i0} = y_0, A_i = a)\phi_\theta^{0|0}(y_2, y_1, y_0) \\
&\qquad\qquad = \frac{e^{\gamma_0 y_0 + a}}{1 + e^{\gamma_0 y_0 + a}}\left(\frac{e^{\gamma_0 + a}}{1 + e^{\gamma_0 + a}}\phi_\theta^{0|0}(1, 1, y_0) + \frac{1}{1 + e^{\gamma_0 + a}}\phi_\theta^{0|0}(0, 1, y_0)\right) \\
&\qquad\qquad\qquad + \frac{1}{1 + e^{\gamma_0 y_0 + a}}\left(\frac{e^{a}}{1 + e^{a}}\phi_\theta^{0|0}(1, 0, y_0) + \frac{1}{1 + e^{a}}\phi_\theta^{0|0}(0, 0, y_0)\right)
\end{aligned}$$

where the second equality uses the logistic hypothesis. By quick inspection, we see that the terms in the first parenthesis have $(1 + e^{\gamma_0 + a})$ in their denominator unlike $\pi^{0|0}(A_i)$. Because $-e^{-\gamma_0}$ is not a *pole* of $\pi^{0|0}(A_i)$, we intuitively conclude that $\phi_\theta^{0|0}(1, 1, y_0) = \phi_\theta^{0|0}(0, 1, y_0) = 0$[5]. This first deduction leaves us with

$$\mathbb{E}\left[\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0}) \mid Y_{i0} = y_0, A_i = a\right] = \frac{1}{1 + e^{\gamma_0 y_0 + a}}\left(\frac{e^{a}}{1 + e^{a}}\phi_\theta^{0|0}(1, 0, y_0) + \frac{1}{1 + e^{a}}\phi_\theta^{0|0}(0, 0, y_0)\right)$$

Now, since $\pi^{0|0}(A_i)$ does not depend on y_0, we must cancel the denominator $(1 + e^{\gamma_0 y_0 + a})$. To achieve this, we must set: $\phi_\theta^{0|0}(1, 0, y_0) = C_0 e^{\gamma y_0}, \phi_\theta^{0|0}(0, 0, y_0) = C_0$ for some constant $C_0 \in \mathbb{R} \setminus \{0\}$. Then,

$$\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{i2}, Y_{i1}, Y_{i0})|Y_{i0} = y_0, A_i = a\right] = C_0\frac{1}{1 + e^{a}}$$

and $C_0 = 1$ is the appropriate normalization to obtain the desired transition function. Of course, the exact same logic applies for $\phi_{\theta_0}^{1|1}(Y_{i2}, Y_{i1}, Y_{i0})$ and $\pi^{1|1}(A_i)$.

This short calculation reveals a useful principle for the general case $T \geq 2$. We learned from the case with two periods that we can search for functions of three

---

[5] A pole of a rational function is a root of its denominator. Formally, we are substituting $u = e^a$ and we are extending $\pi^{0|0}(u)$ to the real line. Then, if $\pi^{0|0}(u) = \mathbb{E}\left[\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0}) \mid Y_{i0} = y_0, U_i = u\right]$, we have $(1 + e^{\gamma_0 y_0}u)(1 + e^{\gamma_0}u)\pi^{0|0}(u) = (1 + e^{\gamma_0 y_0}u)(1 + e^{\gamma_0}u)\mathbb{E}\left[\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0}) \mid Y_{i0} = y_0, U_i = u\right]$. Since $-e^{-\gamma_0}$ is not a pole of $\pi^{0|0}(u)$, evaluating this equality at $u = -e^{-\gamma_0}$ delivers $0 = \phi_\theta^{0|0}(0, 1, y_0) - \phi_\theta^{0|0}(1, 1, y_0) \implies \phi_\theta^{0|0}(0, 1, y_0) = \phi_\theta^{0|0}(1, 1, y_0)$. Next, $\lim_{u\to+\infty}\pi^{0|0}(u) = 0 = \lim_{u\to+\infty}\mathbb{E}\left[\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, Y_{i0}) \mid Y_{i0} = y_0, U_i = u\right] = \phi_\theta^{0|0}(1, 1, y_0)$. This justifies the claim that $\phi_\theta^{0|0}(0, 1, y_0) = \phi_\theta^{0|0}(1, 1, y_0) = 0$.

consecutive outcomes $\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1})$ such that:

$$\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}) = \mathbb{1}\{Y_{it} = k\}\phi_\theta^{k|k}(Y_{it+1}, k, Y_{it-1})$$
$$\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}) \mid Y_{i0}, Y_{i1}^{t-1}, A_i\right] = \pi^{k|k}(A_i)$$

The first restriction is a functional form that eliminates terms with inadequate poles after taking expectations. The second restriction is a normalization condition to match the desired transition probability. Following this argument, we arrive at the expressions in Lemma 1.

Lemma 1. *In model (AR1 pure) with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, let*

$$\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}) = (1 - Y_{it})e^{\gamma Y_{it+1} Y_{it-1}}$$
$$\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}) = Y_{it}e^{\gamma(1-Y_{it+1})(1-Y_{it-1})}$$

Then:

$$\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1})|Y_{i0}, Y_{i1}^{t-1}, A_i\right] = \pi^{0|0}(A_i) = \frac{1}{1 + e^{A_i}}$$
$$\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1})|Y_{i0}, Y_{i1}^{t-1}, A_i\right] = \pi^{1|1}(A_i) = \frac{e^{\gamma_0 + A_i}}{1 + e^{\gamma_0 + A_i}}$$

Step 2). The second step in the agenda is the construction of valid moment functions. By virtue of the law of iterated expectations and since the transition probabilities of the model are time-invariant, a natural way to achieve this is to consider the pairwise difference of $\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1})$ and $\phi_\theta^{k|k}(Y_{is+1}, Y_{is}, Y_{is-1})$ for any feasible $s \neq t$. Nevertheless, alternative differencing schemes are possible and we formally discuss one that can further accommodate arbitrary regressors in Proposition 1 below.

3.2.2 The general case with regressors

We move on to the general AR(1) logit model characterized by equation (AR1).

Step 1). Since the transition probabilities $\pi_t^{0|0}(A_i, X_i), \pi_t^{1|1}(A_i, X_i)$ retain the same functional form as in the simple pure model, the same calculations described above lead to the transition functions in Lemma 2. The only predictable change is the appearance of an extra term $+/-\Delta X_{it+1}'\beta$ which accounts for the presence of covariates in the model.

Lemma 2. *In model (AR1) with $T \geq 2$ and $t \in \{1, \dots, T-1\}$, let*

$$\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = (1 - Y_{it}) e^{Y_{it+1}\left(\gamma Y_{it-1} - \Delta X'_{it+1}\beta\right)}$$
$$\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = Y_{it} e^{(1-Y_{it+1})\left(\gamma(1-Y_{it-1}) + \Delta X'_{it+1}\beta\right)}$$

Then:

$$\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{0|0}(A_i, X_i) = \frac{1}{1 + e^{A_i + X'_{it+1}\beta_0}}$$
$$\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{1|1}(A_i, X_i) = \frac{e^{\gamma_0 + X'_{it+1}\beta_0 + A_i}}{1 + e^{\gamma_0 + X'_{it+1}\beta_0 + A_i}}$$

At this point, it is important to highlight that unlike previously, the transition probabilities are covariate-dependent. As a consequence, the naive difference of $\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)$ and $\phi_\theta^{k|k}(Y_{is+1}, Y_{is}, Y_{is-1}, X_i)$ for $s \neq t$ no longer leads to valid moment functions in general. Indeed, while Lemma 2 ensures that

$$\mathbb{E}\left[\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) - \phi_\theta^{k|k}(Y_{is+1}, Y_{is}, Y_{is-1}, X_i)|Y_{i0}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i) - \pi_s^{k|k}(A_i, X_i)$$

clearly, $\pi_t^{k|k}(A_i, X_i) - \pi_s^{k|k}(A_i, X_i) \neq 0$ when $X'_{it+1}\beta_0 \neq X'_{is+1}\beta_0$ [6]. Thus, a different logic is required in the presence of explanatory variables other than a first order lag. Our proposal is to construct new transition functions that we denote ζ_θ, distinct from $\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)$ but mapping to the same transition probabilities $\pi_t^{k|k}(A_i, X_i)$. To this end, the key idea is to exploit the partial fraction decomposition (see Appendix Lemma 6)

$$(1 - e^{u-v})\frac{e^{v+a}}{(1 + e^{v+a})(1 + e^{u+a})} = \frac{1}{1 + e^{u+a}} - \frac{1}{1 + e^{v+a}} \tag{3}$$

where a plays the role of the fixed effect while u, v capture covariates in different time periods. This algebraic identity shows that a certain weighted product of two distinct rational fractions in $\exp(a)$ can be rewritten as their difference; a standard manipulation to integrate rational functions in real analysis. In our context, this insight motivates considering quantities such as $\mathbb{E}\left[\mathbb{1}\{Y_{is} \neq k\}\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, X_i, A_i\right]$, $s < t$, whose functional form mirrors the left-hand side of (3). After suitable rescaling, (3) indicates that it will coincide with the difference $\mathbb{E}\left[\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, X_i, A_i\right] - \mathbb{E}\left[\mathbb{1}\{Y_{is} = k\}|Y_{i0}, X_i, A_i\right]$,

[6] A matching strategy a la Honoré and Kyriazidou (2000) may still be applicable if $X_{it+1} = X_{is+1}$. However, this is known to lead to estimators converging at rate less than $\sqrt{N}$ for continuous covariates and it rules out certain regressors such as time dummies and time trends.

allowing us to back out the form of the transition functions ζ_θ displayed in Lemma 3. The lemma slightly generalizes this reasoning and formalizes the construction when $T \geq 3$.

Lemma 3. *In model (AR1) with $T \geq 3$, for all t, s such that $T-1 \geq t > s \geq 1$, let:*

$$\begin{aligned}
\mu_s(\theta) &= \gamma Y_{is-1} + X_{is}'\beta \\
\kappa_t^{0|0}(\theta) &= X_{it+1}'\beta, \quad \kappa_t^{1|1}(\theta) = \gamma + X_{it+1}'\beta \\
\omega_{t,s}^{0|0}(\theta) &= 1 - e^{(\kappa_t^{0|0}(\theta) - \mu_s(\theta))}, \quad \omega_{t,s}^{1|1}(\theta) = 1 - e^{-(\kappa_t^{1|1}(\theta) - \mu_s(\theta))}
\end{aligned}$$

and define the moment functions:

$$\begin{aligned}
\zeta_\theta^{0|0}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) &= (1 - Y_{is}) + \omega_{t,s}^{0|0}(\theta) Y_{is} \phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) \\
\zeta_\theta^{1|1}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) &= Y_{is} + \omega_{t,s}^{1|1}(\theta)(1 - Y_{is}) \phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)
\end{aligned}$$

Additionally, if $T \geq 4$, for any t and ordered collection of indices s_1^J, $J \geq 2$, satisfying $T-1 \geq t > s_1 > \ldots > s_J \geq 1$, define analogously

$$\begin{aligned}
\zeta_\theta^{0|0}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) &= (1 - Y_{is_J}) + \omega_{t,s_J}^{0|0}(\theta) Y_{is_J} \zeta_\theta^{0|0}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_{J-1}-1}^{s_{J-1}}, X_i) \\
\zeta_\theta^{1|1}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) &= Y_{is_J} + \omega_{t,s_J}^{1|1}(\theta)(1 - Y_{is_J}) \zeta_\theta^{1|1}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_{J-1}-1}^{s_{J-1}}, X_i)
\end{aligned}$$

Then for all $k \in \mathcal{Y}$

$$\begin{aligned}
\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) | Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] &= \pi_t^{k|k}(A_i, X_i) \\
\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) | Y_{i0}, Y_{i1}^{s_J-1}, X_i, A_i\right] &= \pi_t^{k|k}(A_i, X_i),
\end{aligned}$$

Stepping back, two ingredients are key for our construction: i) the rational fraction structure of the transition probabilities with respect to $\exp(A_i)$, and ii) the availability of suitable partial fraction decompositions (Appendix Lemma 6). The latter relate to the hyperbolic transformations ideas of Kitazawa (2022). In the sequel, we shall see that thoses insights carry over to other DFEL models, including AR(p) logit models for arbitrary $p \geq 1$. **Step 2)**. Provided $T \geq 3$, the difference between any transition functions associated to the same transition probabilities in periods $t \in \{2, \ldots, T-1\}$ constitutes a valid candidate for (1) by iterated expectations. Proposition 1 gives a particular family of valid moment functions that we have found to be complete.

Proposition 1. *In model (AR1), for all $k \in \mathcal{Y}$,*
if $T \geq 3$, for all t, s such that $T-1 \geq t > s \geq 1$, let

$$\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) = \phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i) - \zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i),$$

if $T \geq 4$, for any t and ordered collection of indices s_1^J, $J \geq 2$, satisfying $T-1 \geq t > s_1 > \ldots > s_J \geq 1$, let

$$\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i) = \phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i) - \zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i),$$

Then,

$$\mathbb{E}\left[\psi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] = 0$$
$$\mathbb{E}\left[\psi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i)|Y_{i0}, Y_{i1}^{s_J-1}, X_i, A_i\right] = 0$$

Indeed, note first that this family has cardinality $2^T - 2T$ which by Theorem 1 is precisely the number of linearly independent moment conditions available for the AR(1). To see this, notice that for fixed $(k, Y_{i0}) \in \mathcal{Y}^2$, and a given time period $t \in \{2, \ldots, T-1\}$, Proposition 1 gives a total of: $\sum_{l=1}^{t-1}\binom{t-1}{l} = 2^{t-1}-1$ valid moment functions. Indeed, we get $\binom{t-1}{1}$ possibilities from choosing any s in $\{1, \ldots, t-1\}$ to form $\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)$. To that, we must add another $\sum_{l=2}^{t-1}\binom{t-1}{l}$ possibilities from choosing all feasible sequences s_1^J with $t-1 \geq s_1 > s_2 > \ldots > s_J \geq 1$ to form $\psi_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i)$. Summing over $t = 2, \ldots, T-1$ and multiplying by 2 to account for the two possible values for k delivers the result: $2 \times \sum_{t=2}^{T-1}\sum_{l=1}^{t-1}\binom{t-1}{l} = 2 \times \sum_{t=2}^{T-1}(2^{t-1}-1) = 2^T - 2T$. Second, the family appears linearly independent. It is readily verified for $T = 3$ since the two valid moment functions produced depend on distinct sets of choice histories. Unfortunately, this argument does not carry over to longer panels, but we verified numerically that the linear independence property of this family continues to hold for several different values of $T \geq 4$. This evidence suggests that our two-step approach delivers all the moment equality restrictions available in the AR(1) logit model.[7]

Remark 2 (Static logit)**.** If $\gamma_0 = 0$, model (AR1) specializes to the static panel logit model of Rasch (1960). For that case, Lemma 2 gives two moment functions for

[7] This is not all the identifying content of the AR(1) specification since we know from Dobronyi et al. (2021) that the model also implies moment inequality conditions.

$T = 2$,

$$\phi_\theta^{0|0}(Y_{i2}, Y_{i1}, X_i) = (1 - Y_{i1})e^{-Y_{i2}\Delta X_2'\beta}$$
$$\phi_\theta^{1|1}(Y_{i2}, Y_{i1}, X_i) = Y_{i1}e^{(1-Y_{i2})\Delta X_{i2}'\beta}$$

such that $\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{i1}^2, , X_i)|X_i, A_i\right] = \frac{1}{1+e^{X_{i2}'\beta_0+A_i}}$ and $\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{i1}^2, X_i)|X_i, A_i\right] = \frac{e^{X_{i2}'\beta_0+A_i}}{1+e^{X_{i2}'\beta_0+A_i}}$. It follows that a valid moment function with two periods of observation is

$$\begin{aligned}\psi_\theta(Y_{i2}, Y_{i1}, X_i) &= \phi_\theta^{1|1}(Y_{i2}, Y_{i1}, X_i) - (1 - \phi_\theta^{0|0}(Y_{i2}, Y_{i1}, X_i)) \\ &= (1 - e^{-\Delta X_{i2}'\beta})\left(Y_{i1}(1 - Y_{i2})e^{\Delta X_{i2}'\beta} - (1 - Y_{i1})Y_{i2}\right)\end{aligned}$$

which is proportional to the score of the conditional likelihood based on the sufficient statistic $Y_{i1} + Y_{i2}$ (Rasch (1960), Andersen (1970), Chamberlain (1980)).

3.3 Semiparametric efficiency bound for the AR(1)

Honoré and Weidner (2020) gave sufficient conditions to identify $\theta_0 = (\gamma_0, \beta_0')'$ in the AR(1) model with $T \geq 3$. Two natural follow-up questions arise: i) how accurately can θ_0 be estimated in that case, i.e what is the semiparametric efficiency bound, and ii) which estimator, if any, attains it. This section addresses these questions which to our knowledge have remained unresolved, particularly in the case where covariates are present.

No covariates with $\mathbf{T = 3}$. In a corrigendum to Hahn (2001), Gu et al. (2023) confirmed that the conditional likelihood estimator is semiparametrically efficient for $T = 3$ in the pure AR(1) model. This result, when viewed through our moment-based framework, reveals useful insights. Specifically, with some algebra, one can show that the conditional score for the state dependence parameter $\theta_0 \equiv \gamma_0$ is given by

$$\frac{1}{(1 + e^{\gamma_0})(e^{-\gamma_0} - 1)}\left(\psi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}) + \psi_{\theta_0}^{1|1}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0})\right)$$

where $\psi_\theta^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0})$ and $\psi_\theta^{1|1}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0})$ are the moment functions of our Proposition 1 for the no-regressor case. This expression implies an alternative interpretation of the optimal estimator as the efficient GMM estimator for $\mathbb{E}\left[\psi_{\theta_0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0})|Y_{i0}\right] = 0$, where $\psi_\theta(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}) = (\psi_\theta(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}), \psi_\theta(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}))'$.

The case with covariates and arbitrary T. The pure AR(1) model insights naturally suggest that the efficient GMM estimator for the conditional moment restriction $\mathbb{E}\left[\psi_\theta(Y_{i0}, Y_i, X_i)|Y_{i0}, X_i\right] = 0$ could achieve semiparametric efficiency. Here, $\psi_\theta(Y_{i0}, Y_i, X_i)$ represents the $(2^T - 2T)$-vector gathering all the valid moment functions of Proposition 1[8]. We verify this conjecture in Theorem 2 below. To set out the result, assume θ_0 is identified from $\mathbb{E}\left[\psi_{\theta_0}(Y_{i0}, Y_i, X_i)|Y_{i0}, X_i\right] = 0$ and let $D(Y_{i0}, X_i) = \mathbb{E}\left[\frac{\partial \psi_{\theta_0}(Y_{i0}, Y_i, X_i)}{\partial \theta'}|Y_{i0}, X_i\right]$ and $\Sigma(Y_{i0}, X_i) = \mathbb{E}\left[\psi_{\theta_0}(Y_{i0}, Y_i, X_i)\psi_{\theta_0}(Y_{i0}, Y_i, X_i)'|Y_{i0}, X_i\right]$. Then we have the following result:

Theorem 2. *Consider model (AR1) with $T \geq 3$ and suppose i) $\mathbb{E}\left[X_i X_i'\right] < \infty$, ii) the support $\mathcal{A}_q \subseteq \mathbb{R}$ of the distribution of heterogeneity $q(.|Y_{i0}, X_i)$ contains an accumulation point, iii) the matrix $\mathbb{E}\left[D(Y_{i0}, X_i)'\Sigma(Y_{i0}, X_i)^{-1}D(Y_{i0}, X_i)\right]$ exists and is nonsingular. Then, the semiparametric variance bound for θ_0 is finite and given by $V_0 = \mathbb{E}[D(Y_{i0}, X_i)'\Sigma(Y_{i0}, X_i)^{-1}D(Y_{i0}, X_i)]^{-1}$.*

Assumption i) is a standard square integrability condition for covariates. Assumption ii) is a richness condition weaker than requiring $\mathcal{A}_q = \mathbb{R}$ but sufficient to ensure that no additional information can come from exploiting the support of heterogeneity (see Argañaraz and Escanciano (2023)). Assumption iii) is a local identification condition analogous to Davezies et al. (2023) in the context of static models. Theorem 2 confirms that optimal GMM estimation of θ_0 would utilize the efficient moment function $\psi_\theta^{eff}(Y_{i0}, Y_i, X_i) = D(Y_{i0}, X_i)'\Sigma(Y_{i0}, X_i)^{-1}\psi_\theta(Y_{i0}, Y_i, X_i)$. Its proof involves verifying the conditions for an application of Theorem 3.2 in Newey (1990) and hinges on two key properties. First, we show that the orthocomplement of the nonparametric tanget set - the space onto which the score for θ is projected to determine the element characterizing the variance bound, i.e the *efficient score* - is the set of valid moment conditions verifying (1) (up to terms in (Y_{i0}, X_i)). Second, we leverage the fact that the AR(1) only admits a known finite number of linearly independent moment restrictions by Theorem 1. Together, these features imply that the efficient score is the conditional linear predictor of the score for θ on $\psi_\theta(Y_{i0}, Y_i, X_i)$ given (Y_{i0}, X_i), aligning with $\psi_\theta^{eff}(Y_{i0}, Y_i, X_i)$. We note that these properties are not unique to AR(1) logit model; they hold, for instance, in AR(p) logit models with $p > 1$ (see Theorem 3). This suggests that Theorem 2 could, in principle, be extended to other DFEL models where θ_0 is identified by the available moment conditions.

[8] More generally, any family of $2^T - 2T$ linearly independent valid moment functions could be used.

4 The AR(p) logit model with $p > 1$

Allowing for higher-order lags is often desirable in empirical work to model persistent stochastic processes and improve model fit (e.g, Magnac (2000) on labour market histories, Chay et al. (1999) and Card and Hyslop (2005) on welfare recipiency). In this section, we characterize the form of the moment restrictions available in AR(p) logit models

$$Y_{it} = \mathbb{1}\left\{\sum_{r=1}^{p} \gamma_{0r} Y_{it-r} + X_{it}'\beta_0 + A_i - \epsilon_{it} \geq 0\right\}, \quad t = 1, \ldots, T \tag{ARp}$$

where the lag order $p \geq 1$ can be arbitrary. This generalization has not been thoroughly addressed in the literature [9] and allows to test lag misspecification given enough time periods. Here, $Y_i^0 = (Y_{i-(p-1)}, \ldots, Y_{i-1}, Y_{i0})' \in \mathcal{Y}^p$, $\mathcal{X} \subseteq \mathbb{R}^{K_x}$, $\theta_0 = (\gamma_0', \beta_0')' \in \mathbb{R}^p \times \mathbb{R}^{K_x}$, and $\mathcal{A} = \mathbb{R}$. The logistic assumption on ϵ_{it} implies 2^p non-redundant transition probabilities given by

$$\pi_t^{k|k_1^p}(A_i, X_i) = P(Y_{it+1} = k_1 | Y_{it} = k_1, \ldots, Y_{it-(p-1)} = k_p, X_i, A_i) = \frac{e^{k_1(\sum_{r=1}^{p}\gamma_{0r}k_r + X_{it+1}'\beta_0 + A_i)}}{1 + e^{\sum_{r=1}^{p}\gamma_{0r}k_r + X_{it+1}'\beta_0 + A_i}}$$

for $(k_1, \ldots, k_p) \in \mathcal{Y}^p$.

4.1 The number of moment restrictions when $p \geq 1$

Based on simulation evidence, Honoré and Weidner (2020) conjectured that AR(p) models possess $2^T - (T + p - 1)2^p$ linearly independent moment conditions in panels of sufficient length. We prove this claim in Theorem 3 and establish that no moment restrictions for the common parameters exist when $T \leq p+1$. To introduce the result formally, it is again convenient to consider the conditional expectation operator $\mathcal{E}_{y^0,x,T}^{(p)}$ describing the model, i.e

$$\mathcal{E}_{y^0,x,T}^{(p)}\left[\mathbb{1}\{. = y\}\right] = P(Y_i = y | Y_i^0 = y^0, X_i = x, A_i = .) = a \mapsto \prod_{t=1}^{T} \frac{e^{y_t\left(\sum_{r=1}^{p}\gamma_{0r}y_{t-r} + x_t'\beta_0 + a\right)}}{1 + e^{\sum_{r=1}^{p}\gamma_{0r}y_{t-r} + x_t'\beta_0 + a}}$$

Then the following result holds:

[9] Using Mathematica, Honoré and Weidner (2020) present moment functions for the AR(2) model up to $T = 4$ and the AR(3) model with $T = 5$ but no results are offered beyond these special cases.

Theorem 3. *Consider model (ARp) with $T \geq 1$, initial condition $y^0 \in \mathcal{Y}^p$ and covariates $x \in \mathcal{X}^T$. Suppose that for any $t, s \in \{1, \ldots, T-1\}$ and $y, \tilde{y} \in \mathcal{Y}^p$, $\gamma_0' y + x_t'\beta_0 \neq \gamma_0'\tilde{y} + x_s'\beta_0$ if $t \neq s$ or $y \neq \tilde{y}$. Then, the family*

$$\mathcal{F}^{(p)}_{y^0,x,T} = \left\{ 1, \pi_0^{y_0|y^0}(.,x), \left\{ \left(\pi_{t-1}^{y_1|y_1^{t-1}, y_0, \ldots, y_{-(p-t)}}(.,x)) \right)_{y_1^{t-1} \in \mathcal{Y}^{t-1}} \right\}_{t=2}^{p}, \left\{ \left(\pi_{t-1}^{y_1|y_1^p}(.,x) \right)_{y_1^p \in \mathcal{Y}^p} \right\}_{t=p+1}^{T} \right\}$$

forms a basis of $\operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)$ *and therefore*

1. *If $T \leq p+1$,* $\operatorname{rank}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right) = 2^T$ *and* $\dim\left(\ker\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)\right) = 0$
2. *If $T \geq p+2$,* $\operatorname{rank}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right) = (T-p+1)2^p$ *and* $\dim\left(\ker\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)\right) = 2^T - (T-p+1)2^p$

Theorem 3 generalizes Theorem 1, establishing that the conditional probabilities of all choice histories are spanned by the transition probabilities, no matter the lag order. This result hinges again on the rational fraction structure of logit probabilities and on the fact that the transition probabilities of AR(p) models admit transition functions, a property set out in the following section. One important practical implication is that fitting an AR(p) demands at least $2(p-1)$ additional observations relative to an AR(1) (count p initial conditions followed by $T = p+2$ waves of data against 4 total periods needed for an AR(1)).

Remark 3 (Beyond Logit)**.** Theorem 1 and 3 could, in principle, be suitably extended to other distributions for ϵ_{it} beyond the logistic case, provided they induce a rational fraction structure for the transition probabilities. Examples include mixtures of logistic distributions (e.g Honoré and Weidner (2020)), and generalized logistic distributions (e.g Davezies et al. (2023)). A rational fraction structure prevents the rank of the conditional expectation operator from growing as quickly as the number of choice histories, ensuring thereby the existence of moment conditions for sufficiently large T.

4.2 Construction of transition functions with $p > 1$

Having clarified that $T = p+2$ is the minimum number of periods required for the existence of identifying moments, we are now ready to address the issue of their construction. The blueprint generalizes that of the AR(1) model and can be summarized as follows:

1. **Step 1)**
   (a) Start by obtaining analytical expressions of the unique transition functions for the transition probability in period $t = p$ when $T = p + 1$ [10]. Shift these expressions by one period, two periods, three periods etc to get a set of transition functions for period $t \in \{p+1, \ldots, T-1\}$ when $T \geq p+2$.
   (b) Apply partial fraction decompositions to the expressions obtained in (a) for $t \in \{p+1, \ldots, T-1\}$ to generate other transition functions mapping to the same transition probabilities.
2. **Step 2)**. Take adequate differences of transition functions associated to the same transition probability in periods $t \in \{p+1, \ldots, T-1\}$ to obtain valid moments that are linearly independent.

Step 1) (a) is akin to how we started by getting closed form expressions for the transition functions in period $t = 1$ for $T = 2$ in the one lag case and then deducted a general principle for $t \geq 2$ (see Section 3.2.1). From a technical perspective, this is the only part of the two-step procedure that differs from the baseline AR(1). Indeed, **Step 2)** is fundamentally identical and **Step 1)** (b) is also unchanged for the simple reason that the transition probabilities keep the same functional form as before. That is, a rational fraction in $\exp(A_i)$. Hence, the same partial fraction expansions outlined in Section 3.2.2 apply. In light of those close similarities with the AR(1) and in order to focus on the primary issues, we defer a discussion of **Step 1)**(b) and **Step 2)** to the Supplemental Appendix.

Theorem 4 provides the algorithm to compute the transition functions for **Step 1)** (a) for arbitrary lag order greater than one. It is based on the insight that we can leverage the transition functions of an AR($p-1$) and partial fraction decompositions to generate the transition functions of an AR(p). A simple example is helpful to illustrate the idea. Consider an AR(2) with $T = 3$ (i.e 5 observations in total) and suppose that we seek a transition function associated to, say, the transition probability $\pi_2^{0|0,1}(A_i, X_i) = \left(1 + e^{\gamma_{02} + X_{i3}'\beta_0 + A_i}\right)^{-1}$.

The first ingredient of the theorem is to view the AR(2) model as an AR(1) model where we treat the second order lag as an additional strictly exogenous regressor. This change of perspective is advantageous since we already know how to deal

[10] The fact that the transition functions in period $t = p$ are unique when $T = p + 1$ is a direct corollary of Theorem 3. Otherwise, the difference of two distinct transition functions mapping to the same transition probability would yield a valid moment which is a contradiction.

with the single lag case. In particular, Lemma 2 readily gives the transition function $\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)$ for the transition probability $\pi_2^{0|0,Y_{i1}}(A_i, X_i) = P(Y_{i3} = 0|Y_{i2} = 0, Y_{i1}, X_i, A_i)$ in the sense that it verifies:

$$\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)|Y_i^0, Y_{i1}, X_i, A_i\right] = \pi_2^{0|0,Y_{i1}}(A_i, X_i)$$

This is an intermediate stage since $\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)$ does not quite map to the target of interest; $\pi_2^{0|0,Y_{i1}}(A_i, X_i)$ depends on the random variable Y_{i1} unlike $\pi_2^{0|0,1}(A_i, X_i)$. To make further progress, one would intuitively need to "set" Y_{i1} to unity to make the two transition probabilities coincide. We operationalize this idea by interacting $\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)$ and Y_{i1} to achieve the desired effect in expectation:

$$\begin{aligned}\mathbb{E}\left[Y_{i1}\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)|Y_i^0, X_i, A_i\right] &= \mathbb{E}\left[Y_{i1}\pi_2^{0|0,1}(A_i, X_i)|Y_i^0, X_i, A_i\right] \\ &= \frac{1}{1 + e^{\gamma_{02} + X_{i3}'\beta + A_i}} \frac{e^{\gamma_{01}Y_{i0} + \gamma_{02}Y_{i-1} + X_{i1}'\beta_0 + A_i}}{1 + e^{\gamma_{01}Y_{i0} + \gamma_{02}Y_{i-1} + X_{i1}'\beta_0 + A_i}}\end{aligned}$$

Here, the first equality follows from the law of iterated expectations. Then, the second ingredient of the theorem is again the partial fraction expansion (3) to turn this product of logistic indices into $\pi_2^{0|0,1}(A_i, X_i)$. This last operation is analogous to how we constructed sequences of transition functions in the AR(1) model. It ultimately tells us that the solution is a weighted sum of $(1-Y_{i1})$ and $Y_{i1}\phi_{\theta_0}^{0|0}(Y_{i3}, Y_{i2}, Y_{i1}, Y_{i0}, X_i)$. Theorem 4 turns this procedure into a recursive algorithm that computes the transition functions for any lag order $p > 1$.

Theorem 4. *In model (ARp) with $T \geq p+1$, for all $t \in \{p, \dots, T-1\}$ and $y_1^p \in \mathcal{Y}^p$, let*

$$k_t^{y_1|y_1^p}(\theta) = \sum_{r=1}^{p} \gamma_r y_r + X_{it+1}'\beta$$

$$k_t^{y_1|y_1^{k+1}}(\theta) = \sum_{r=1}^{k+1} \gamma_r y_r + \sum_{r=k+2}^{p} \gamma_r Y_{it-(r-1)} + X_{it+1}'\beta, \quad k = 1, \dots, p-2, \text{ if } p > 2$$

$$u_{t-k}(\theta) = \sum_{r=1}^{p} \gamma_r Y_{it-(r+k)} + X_{it-k}'\beta, \quad k = 1, \dots, p-1$$

$$w_t^{y_1|y_1^{k+1}}(\theta) = \left[1 - e^{(k_t^{y_1|y_1^{k+1}}(\theta) - u_{t-k}(\theta))}\right]^{y_{k+1}} \left[1 - e^{-(k_t^{y_1|y_1^{k+1}}(\theta) - u_{t-k}(\theta))}\right]^{1-y_{k+1}}, \quad k = 1, \dots, p-1$$

and

$$\phi_{\theta}^{y_1|y_1^{k+1}}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i) =$$
$$\left[(1-Y_{it-k}) + w_t^{y_1|y_1^{k+1}}(\theta)\phi_{\theta}^{y_1|y_1^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)Y_{it-k}\right]^{(1-y_1)y_{k+1}} \times$$
$$\left[1 - Y_{it-k} - w_t^{y_1|y_1^{k+1}}(\theta)\left(1 - \phi_{\theta}^{y_1|y_1^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)\right)(1-Y_{it-k})\right]^{(1-y_1)(1-y_{k+1})} \times$$
$$\left[Y_{it-k} + w_t^{y_1|y_1^{k+1}}(\theta)\phi_{\theta}^{y_1|y_1^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)(1-Y_{it-k})\right]^{y_1(1-y_{k+1})} \times$$
$$\left[1 - (1-Y_{it-k}) - w_t^{y_1|y_1^{k+1}}(\theta)\left(1 - \phi_{\theta}^{y_1|y_1^k}(Y_{it+1}, Y_{it}, Y_{it-(p+k-1)}^{t-1}, X_i)\right)Y_{it-k}\right]^{y_1 y_{k+1}}, \quad k = 1, \ldots, p-1$$

where

$$\phi_{\theta}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) = (1-Y_{it})e^{Y_{it+1}(\gamma_1 Y_{it-1} - \sum_{l=2}^{p}\gamma_l \Delta Y_{it+1-l} - \Delta X_{it+1}'\beta)}$$
$$\phi_{\theta}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-p}^{t-1}, X_i) = Y_{it}e^{(1-Y_{it+1})\left(\gamma_1(1-Y_{it-1}) + \sum_{l=2}^{p}\gamma_l \Delta Y_{it+1-l} + \Delta X_{it+1}'\beta\right)}$$

Then,

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^p}(Y_{it+1}, Y_{it}, Y_{it-(2p-1)}^{t-1}, X_i) \mid Y_i^0, Y_{i1}^{t-p}, X_i, A_i\right] = \pi_t^{y_1|y_1^p}(A_i, X_i)$$

and for $k = 0, \ldots, p-2$

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{k+1}}(Y_{it+1}, Y_{it}, Y_{it-(p+k)}^{t-1}, X_i) \mid Y_i^0, Y_{i1}^{t-(k+1)}, X_i, A_i\right] = \pi_t^{y_1|y_1^{k+1}, Y_{it-(k+1)}, \ldots, Y_{it-(p-1)}}(A_i, X_i)$$

The remaining steps to complete the construction of valid moment functions are described at length in the Supplemental Appendix. The end product is a family of (numerically) linearly independent moment functions of size $2^T - (T+1-p)2^p$. By Theorem 3, this implies that our two-step approach recovers all moment equality conditions in the model. We discuss how to potentially exploit these moment functions to identify θ_0 in the Supplemental Appendix.

Remark 4 (Extensions)**.** While the exposition emphasized model (ARp), the methodology applies more broadly to models of the form $Y_{it} = \mathbb{1}\left\{g(Y_{it-1}, \ldots, Y_{it-p}, X_{it}, \theta_0) + A_i - \epsilon_{it} \geq 0\right\}$, where the lag order $p > 1$ is known and $g(.)$ is known up to θ_0. The crucial feature is the additive separability of the fixed effect.

Remark 5. (Average Marginal Effects) In applied work, there is often interest in certain functionals of unobserved heterogeneity rather than on the value of the model parameters per se. Average marginal effects (AMEs) which capture mean response to a counterfactual change in past outcomes are one such example, and can be directly obtained as expectations of our transition functions. To illustrate, consider the baseline AR(1) model with discrete covariates X_{it}. We can define the average transition probability from state l to state k in period t for a subpopulation of individuals with covariate $x_1^{t+1} = (x_1, \ldots, x_{t+1})$ and initial condition y_0 as

$$\Pi_t^{k|l}(y_0, x_1^{t+1}) = \mathbb{E}\left[\underbrace{\pi_t^{k|l}(X_{it+1}, A_i)}_{\equiv \pi_t^{k|l}(X_i, A_i)} \,|\, Y_{i0} = y_0, X_{i1}^{t+1} = x_1^{t+1}\right] = \int \pi_t^{k|l}(x_{t+1}, a) q(a|y_0, x_1^{t+1}) da$$

where $q(.|y_0, x_1^{t+1})$ denotes the conditional density of the fixed effect given (y_0, x_1^{t+1}). The AME is defined as the following contrast of average transition probabilities:

$$AME_t(y_0, x_1^{t+1}) = \Pi_t^{1|1}(y_0, x_1^{t+1}) - \Pi_t^{1|0}(y_0, x_1^{t+1}) = \Pi_t^{1|1}(y_0, x_1^{t+1}) - (1 - \Pi_t^{0|0}(y_0, x_1^{t+1}))$$

It is interpreted as the population average causal effect on Y_{it+1} of a change from 0 to 1 of Y_{it} given (y_0, x_1^{t+1}). By Lemma 2 and the law of iterated expectations, we have that for $T \geq 2$ and $t \geq 1$: $\Pi_t^{k|k}(y_0, x_1^{t+1}) = \mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) \,|\, Y_{i0} = y_0, X_{i1}^{t+1} = x_1^{t+1}\right], k \in \mathcal{Y}$, implying that $AME_t(y_0, x_1^{t+1})$ is identified so long as θ_0 is identified. A sufficient condition for that is $T \geq 3$ and $X_{i3} - X_{i2}$ having support in a neighborhood of zero (Honoré and Kyriazidou (2000)). Aguirregabiria and Carro (2021) were the first to point out the identification of AMEs in the AR(1) model. Theorem 4 shows that our transition functions can be leveraged more broadly to recover AMEs in AR(p) logit models with $p > 1$. Naturally, this insight extends to any average effect whose integrand can be expressed as a linear combination of transition probabilities. This includes, for example, "average survivor functions", representing counterfactual probabilities of surviving s consecutive periods in the same state.

5 Moment restrictions for the VAR(1) logit model

We now turn our attention to multi-dimensional fixed effects models, focusing in this section on the VAR(1) logit used in our empirical application. Readers will find the proofs of all claims in this section and analogous results for the dynamic multinomial logit model in the Supplemental Appendix.

Let $Y_{it} = (Y_{1,it}, \ldots, Y_{M,it})' \in \mathcal{Y} = \{0,1\}^M$ denote the outcome vector in period t with $M \geq 2$. Let $X_{it} = (X'_{1,it}, \ldots, X'_{M,it})' \in \mathcal{X} \subseteq \mathbb{R}^{K_1} \times \ldots \times \mathbb{R}^{K_M}$ denote the vector of exogenous covariates in period t and $A_i = (A_{1,i}, \ldots, A_{M,i})' \in \mathcal{A} = \mathbb{R}^M$ the vector of fixed effects. The VAR(1) logit model is described by:

$$Y_{m,it} = \mathbb{1}\left\{\sum_{j=1}^{M} \gamma_{0mj} Y_{j,it-1} + X'_{m,it}\beta_{0m} + A_{m,i} - \epsilon_{m,it} \geq 0\right\} \tag{VAR1}$$

$m = 1, \ldots, M, \quad t = 1, \ldots, T$. It represents a natural extension of the baseline AR(1) logit model for multivariate outcomes and has been applied to study the relationship between sickness and unemployment (Narendranthan et al. (1985)), the progression from softer drug use to harder drug use among teenagers (Deza (2015)), transitivity in networks (Graham (2013), Graham (2016)) and more recently the employment of couples (Honoré et al. (2022)). The initial condition is given by $Y_{i0} = (Y_{1,i0}, \ldots, Y_{M,i0})' \in \mathcal{Y}$, and the logistic assumption induces the transition probabilities:

$$\pi_t^{k|l}(A_i, X_i) = P(Y_{it+1} = k | Y_{it} = l, X_i, A_i) = \prod_{m=1}^{M} \frac{e^{k_m(\sum_{j=1}^{M} \gamma_{0mj} l_j + X'_{m,it+1}\beta_{0m} + A_{m,i})}}{1 + e^{\sum_{j=1}^{M} \gamma_{0mj} l_j + X'_{m,it+1}\beta_{0m} + A_{m,i}}}$$

for all $(k,l) \in \mathcal{Y}^2$. Honoré and Kyriazidou (2019) studied the bivariate case and showed that θ_0 can be identified by a conditional likelihood approach if $T \geq 3$ and the regressors do not vary over the last two periods. Similarly to the AR(1) case, the alternative construction for identifying moments below relaxes these restrictions on the covariates, thus allowing for the inclusion of time effects and estimation of common parameters at $\sqrt{N}$-rate.

Step 1) in the VAR(1) logit model has a nuance relative to its univariate counterpart: according to our calculations, the only transition functions that seem to exist are those associated to $\pi_t^{k|k}(A_i, X_i)$, for $k \in \mathcal{Y}$, i.e the probabilities of remaining in the same state. The expressions of a first set of transition functions, available from $T = 2$, are presented in Lemma 4. They can easily be derived by applying the reasoning outlined in subsection 3.2.1.

Lemma 4. *In model (VAR1) with $T \geq 2$ and $t \in \{1, \ldots, T-1\}$, let for all $k \in \mathcal{Y}$*

$$\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = \mathbb{1}\{Y_{it} = k\} e^{\sum_{m=1}^{M}(Y_{m,it+1} - k_m)\left(\sum_{j=1}^{M} \gamma_{mj}(Y_{j,it-1} - k_j) - \Delta X'_{m,it+1}\beta_m\right)}$$

Then:

$$\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i) = \prod_{m=1}^{M} \frac{e^{k_m(\sum_{j=1}^M \gamma_{0mj}k_j + X'_{m,it+1}\beta_{0m} + A_{m,i})}}{1 + e^{\sum_{j=1}^M \gamma_{0mj}k_j + X'_{m,it+1}\beta_{0m} + A_{m,i}}}$$

Next, we can appeal to the second partial fraction decomposition formula in Appendix Lemma 7 to guide the construction of another set of transition functions when $T \geq 3$. The idea is as usual to utilize the (multivariate) rational fraction structure of the transition probabilities. As is clear from Lemma 5, the resulting transition functions are multivariate analogs of those presented in Lemma 3 for the AR(1) model.

Lemma 5. *In model (VAR1) with $T \geq 3$, for all t, s such that $T-1 \geq t > s \geq 1$, let for all $m \in \{1, \dots, M\}$ and $(k,l) \in \mathcal{Y}^2$: $\mu_{m,s}(\theta) = \sum_{j=1}^M \gamma_{mj} Y_{j,is-1} + X'_{m,is}\beta_m$, $\kappa_{m,t}^{k|k}(\theta) = \sum_{j=1}^M \gamma_{mj}k_j + X'_{m,it+1}\beta_m$, $\omega_{t,s,l}^{k|k}(\theta) = 1 - e^{\sum_{j=1}^M (l_j - k_j)\left[\kappa_{j,t}^{k|k}(\theta) - \mu_{j,s}(\theta)\right]}$ and define the moment functions*

$$\zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i) = \mathbb{1}\{Y_{is} = k\} + \sum_{l \in \mathcal{Y}\setminus\{k\}} \omega_{t,s,l}^{k|k}(\theta)\mathbb{1}\{Y_{is} = l\}\phi_\theta^{k|k}(Y_{it-1}^{t+1}, X_i)$$

Additionally, if $T \geq 4$, for any t and ordered collection of indices s_1^J, $J \geq 2$, satisfying $T-1 \geq t > s_1 > \ldots > s_J \geq 1$, define analogously

$$\begin{aligned}\zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \dots, Y_{is_J-1}^{s_J}, X_i) &= \mathbb{1}\{Y_{is_J} = k\} \\ &+ \sum_{l \in \mathcal{Y}\setminus\{k\}} \omega_{t,s_J,l}^{k|k}(\theta)\mathbb{1}\{Y_{is_J} = l\}\zeta_\theta^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \dots, Y_{is_{J-1}-1}^{s_{J-1}}, X_i)\end{aligned}$$

Then

$$\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i)$$
$$\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \dots, Y_{is_J-1}^{s_J}, X_i)|Y_{i0}, Y_{i1}^{s_J-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i)$$

For **Step 2)**, a family of linearly independent valid moment functions is readily available by adequately repurposing the statement of Proposition 1 to the VAR(1) case, i.e by updating the expressions of $\phi_\theta^{k|k}(.)$ and $\zeta_\theta^{k|k}(.)$ according to Lemmas 4-5. To conserve on space and avoid repetition, we leave this simple exercise to the reader.

Remark 6 (Non-exhaustiveness)**.** Although it can be verified numerically that, for $T = 3$, our two-step strategy based on transition functions accounts for all moment restrictions in both the VAR(1) specification and the dynamic multinomial logit model (see Supplemental Appendix), it no longer holds for $T \geq 4$. One can show that there exists functions of the fixed effects beyond linear combinations of transition probabilities that we can difference out using a broader class of "generalized transition functions". Importantly, the resulting moment conditions contain additional information on θ_0 unlike in the binary case. Characterizing the complete family of moment conditions is a complex problem that we address for the dynamic panel multinomial logit model in Dano et al. (2025).

6 Empirical Illustration

In this section, we apply our methodology to analyze the dynamics of drug consumption among young individuals in the United States. The substantive question is whether the observed persistence in drug use and the progression from soft to hard drugs among youth, as documented in studies such as Deza (2015) [11], stem from causal state dependence (within and between drugs) or from latent traits predisposing individuals to illicit substance use.

To investigate these issues, we employ the the National Longitudinal Survey of Youth 1997 (NLSY97) which is a panel dataset of 8984 individuals surveyed on a diverse range of subjects, including drug-related matters from 1997 to 2021 [12]. We concentrate on a subsample of four waves, spanning from 2001 to 2004. This subsample provides insight into the behavior of young people between the age of 16 and 22 in 2001 to 19 and 25 in 2004. We examine the statistical association between three outcome variables, namely the consumption of alcohol, marijuana and hard drugs, derived from respondents answers' during annual interviews. Upon retaining those providing answers in all four waves, our sample consists of $N = 6461$ individuals. In the spirit of Deza (2015), we model the relationship between the consumption of each substance as a trivariate VAR(1) logit model:

$$Y_{m,it} = \mathbb{1}\left\{\sum_{j=1}^{3}\gamma_{0mj}Y_{j,it-1} + \beta_{0m}age_{it} + \delta_{0m}college_{it} + A_{m,i} - \epsilon_{m,it} \geq 0\right\}$$

[11] To fix ideas, in the NLSY97 dataset, the empirical probability of consuming a substance in year $t+1$ conditional on consuming it in year t averaged over $t = 2001, 2002, 2003$ is: 0.82 for alcohol, 0.6 for marijuana, 0.4 for hard drugs. Likewise, the average empirical probability of consuming hards drugs in $t+1$ conditional on consuming marijuana in t over the same periods is approximately 0.16. In contrast, the average empirical probability of consuming hards drugs in $t+1$ conditional on not consuming marijuana in t is only 0.02.

[12] The views expressed here are those of the author and do not reflect the views of the Bureau of Labor Statistics (BLS).

$m \in \{1,2,3\}$ (1="alcohol", 2="marijuana", 3="hard drugs"), $t = 1,2,3$ where $t = 0$ corresponds to the year 2001. The state-dependence coefficients γ_{0mm} (within) and $\gamma_{0mj}, m \neq j$ (between) are the main coefficients of interest in the 15-dimensional vector of common parameters θ_0. We are particularly concerned about the sign and the statistical significance of γ_{032}, i.e the so-called "stepping-stone" effect of marijuana on hard drugs. The covariate age_{it} denotes the age of respondent i at time t, and $college_{it}$ is a dummy variable indicating enrollment in a college degree. It captures the possibility that college represents a drug-friendly environment[13]. Deza (2015) parameterizes both the latent permanent heterogeneity A_i and the initial condition Y_{i0} to estimate the model by maximum likelihood. We leave these components unrestricted and exploit the valid moment functions presented in Section 5. We specifically use six of the eight valid moment functions available: $\psi_\theta^{k|k}(Y_{i1}^3, Y_{i0}^1, X_i)$ for $k \in \{(0,0,0),(0,1,0),(1,1,1),(1,1,0),(1,0,1),(1,0,0)\}$. The other two corresponding to states $k \in \{(0,0,1),(0,1,1)\}$ are null for over 99.5% of our sample and were dropped to mitigate noise in estimation. Next, we select a constant, the initial condition Y_i^0, age_{it} and $college_{it}$ in all time periods to form the 60×1 moment vector

$$m_\theta(Y_i, Y_i^0, X_i) = \begin{pmatrix} \psi_\theta^{(0,0,0)|(0,0,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(0,1,0)|(0,1,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,1,1)|(1,1,1)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,1,0)|(1,1,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,0,1)|(1,0,1)}(Y_{i1}^3, Y_{i0}^1, X_i) \\ \psi_\theta^{(1,0,0)|(1,0,0)}(Y_{i1}^3, Y_{i0}^1, X_i) \end{pmatrix} \otimes \begin{pmatrix} 1 \\ Y_i^{0\prime} \\ age_{i1}^{3\prime} \\ college_{1,i1}^{3\prime} \end{pmatrix}$$

With $m_\theta(Y_i, Y_i^0, X_i)$ in hand, and given the number of overidentifying restrictions, we then consider the empirical likelihood (EL) estimator $\hat{\theta}$ solution to

$$\max_{\theta,\pi} \sum_{i=1}^N \ln \pi_i \quad \text{subject to} \quad \sum_{i=1}^N \pi_i = 1, \quad \sum_{i=1}^N \pi_i m_\theta(Y_i, Y_i^0, X_i) = 0$$

(Qin and Lawless (1994)), motivated by much work documenting the better small sample properties of EL relative to GMM (e.g Imbens (1997) in a panel context). Notably, Newey and Smith (2004) showed that EL has relatively low asymptotic

[13] An earlier version of this paper examined a similar model, replacing college enrollment with the ratio of state-level admissions to treatment centers for drug m in state i and year t to the national counterpart in the same year, following Deza (2015). The results, comparable to those in Tables 1, showed statistically insignificant effects for these alternative regressors. Moreover, constructing these regressors required access to restricted BLS data, which complicated the analysis and limited replicability without providing additional insights. These challenges motivated the adoption of the slightly different specification considered here, which relies on publicly available data.

bias which does not grow with the number of moment restrictions in contrast to GMM. Also, EL is efficient and avoids arbitrary choices of initial consistent estimator and weight matrix as in 2-step GMM (Imbens (1997)). The downside of EL relative to GMM as is well known is computational, demanding in the above formulation to solve a constrained optimization problem with $N + dim(\theta)$ unknowns compared to an unconstrained problem with $dim(\theta)$ unknowns for GMM. However, this was not an issue for this particular application: solving for $\hat{\theta}$ was a matter of a few minutes using Julia on a modern computer. Under suitable regularity conditions (Newey and Smith (2004)), the EL estimator is normally distributed with: $\sqrt{N}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(M'\Omega^{-1}M\right)^{-1}\right)$, where $M = \mathbb{E}\left[\frac{\partial m_{\theta_0}(Y_i, Y_i^0, X_i)}{\partial \theta'}\right]$ and $\Omega = \mathbb{E}\left[m_{\theta_0}(Y_i, Y_i^0, X_i) m_{\theta_0}(Y_i, Y_i^0, X_i)'\right]$. Efficient estimators of M and Ω are given by $\hat{M} = \sum_{i=1}^{N} \hat{\pi}_i \frac{\partial m_{\hat{\theta}}(Y_i, Y_i^0, X_i)}{\partial \theta'}$ and $\hat{\Omega} = \sum_{i=1}^{N} \hat{\pi}_i m_{\hat{\theta}}(Y_i, Y_i^0, X_i) m_{\hat{\theta}}(Y_i, Y_i^0, X_i)'$ where $\hat{\pi}_i$, $i = 1, \ldots, N$ are the EL probabilities.

Table 1 presents the EL estimates for the trivariate VAR(1) logit model in columns (I), (II), (III). For comparison, columns (IV), (V), (VI) report a random effect (RE) estimator akin to Deza (2015) [14] while columns (VII), (VIII), (IV) display the "naive" logit maximum likelihood estimator (MLE) which fits the same model but neglects the presence of fixed effects. The first observation is that, in line with conventional wisdom, EL estimates for the state-dependence parameters within drug, $\gamma_{11}, \gamma_{22}, \gamma_{33}$, are all positive and statistically significant. There is a sharp contrast in the magnitude of these estimates relative to the other two estimators however. The naive MLE largely overestimates the amount of within state-dependence, yielding coefficients that are comparatively three to five times larger. Intuitively, this may be rationalized by the fact that it misinterprets the serial correlation produced by the fixed effects as evidence of state dependence. The RE estimator acts as an intermediate case between the other two as can be seen in columns (IV)-(VI). This behavior is not unexpected to the extent that RE accounts to some degree for the presence of unobserved heterogeneity. We note nevertheless that the role of within state dependence seems overstated by this approach.

Second, EL estimates in column (III) indicate a positive and statistically significant effect of marijuana on hard drugs, although the standard errors are a bit large. This supports the view that marijuana usage may be a gateway to the consumption of harder drugs and accords with the core findings of Deza (2015). The other

[14] As in Deza (2015), the heterogeneity distribution is discrete with 3 mass points and is independent of the regressors. The initial condition relates to the covariates and heterogeneity through a logistic regression.

two estimators also agree on a positive influence of marijuana on the consumption of harder drugs, albeit it is statistically insignificant in the RE case. Additionally, the more robust EL estimates suggest that alcohol does not play a significant role in the consumption of either drug unlike RE and MLE.

We also computed two overidentification test statistics, presented in the bottom rows of Table 1. The first is the empirical likelihood ratio test $LR = -2\left(\sum_{i=1}^{N} \ln \hat{\pi}_i - \ln \frac{1}{N}\right)$. The second is a variant of the usual overidentification test which uses the efficient weight matrix:

$$Wald = \frac{1}{N}\left(\sum_{i=1}^{N} m_{\hat{\theta}}(Y_i, Y_i^0, X_i)\right)\left(\sum_{i=1}^{N} \hat{\pi}_i m_{\hat{\theta}}(Y_i, Y_i^0, X_i) m_{\hat{\theta}}(Y_i, Y_i^0, X_i)'\right)^{-1}\left(\sum_{i=1}^{N} m_{\hat{\theta}}(Y_i, Y_i^0, X_i)\right)$$

In large samples, $LR, Wald \xrightarrow{d} \chi^2(45)$, where the degrees of freedom correspond to the number of overidentifying restrictions (see, e.g Imbens (1997)). As both test values fall below the 90th quantile of a $\chi^2(45)$, the trivariate VAR(1) logit model appears appropriate. Additional estimates for the iterated GMM estimator of Hansen et al. (1996) are reported in Table 2 of the Supplemental Appendix. The results closely mirror those for EL in Table 1.

Table 1: Parameter estimates of the trivariate VAR(1) logit based on NLSY97 data

	Empirical Likelihood			Random Effects			Naive MLE		
	A (I)	M (II)	HD (III)	A (IV)	M (V)	HD (VI)	A (VII)	M (VIII)	HD (IV)
γ_{m1}	0.48	-0.06	0.38	1.45	-0.39	-0.28	2.47	0.88	0.81
	(0.13)	(0.21)	(0.33)	(0.10)	(0.09)	(0.18)	(0.05)	(0.06)	(0.10)
γ_{m2}	0.29	0.83	0.49	-0.49	1.44	0.08	0.70	2.56	1.41
	(0.20)	(0.14)	(0.24)	(0.09)	(0.09)	(0.11)	(0.06)	(0.05)	(0.08)
γ_{m3}	-0.29	0.19	0.48	-0.59	-0.20	1.60	0.25	0.72	2.11
	(0.31)	(0.22)	(0.23)	(0.18)	(0.12)	(0.10)	(0.12)	(0.08)	(0.09)
age	0.09	-0.09	0.03	0.16	-0.14	-0.07	-0.04	-0.14	-0.21
	(0.05)	(0.07)	(0.10)	(0.02)	(0.02)	(0.03)	(0.00)	(0.00)	(0.00)
college	0.25	0.20	0.31	0.75	-0.05	-0.20	0.42	-0.05	-0.24
	(0.14)	(0.15)	(0.26)	(0.06)	(0.06)	(0.08)	(0.04)	(0.04)	(0.07)
LR Test	56.45								
"Wald" Test	54.38								

Notes: standard errors are reported in parenthesis. Columns titled "A","M", "HD" report parameter estimates for the alcohol layer, marijuana layer, and hard-drugs layer of the trivariate VAR(1) logit model.

7 Conclusion

Dynamic discrete choice models are widely used to study the determinants of repeated decisions made by economic agents over time. This paper has introduced a systematic procedure to estimate a large class of such models with logistic (or Type I extreme value) errors and potentially many lags, all while remaining agnostic to unobserved individual heterogeneity. Our application underscores the practical value of the methodology.

There are several interesting directions for future research. One natural question is whether the tools developed here can be deployed in other discrete choice frameworks with similar or even more flexible structure. Another challenge lies in deriving complete basis of moment restrictions beyond the binary response case for arbitrary time horizons. We are investigating some of these topics in ongoing work.

References

Aguirregabiria, V. and Carro, J. M. (2021). Identification of average marginal effects in fixed effects dynamic discrete choice models. *arXiv preprint arXiv:2107.06141*.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 32(2):283–301.

Argañaraz, F. and Escanciano, J. C. (2023). On the existence and information of orthogonal moments. *arXiv preprint arXiv:2303.11418*.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer.

Bonhomme, S. (2012). Functional differencing. *Econometrica*, 80(4):1337 – 1385.

Bradley, W. T. and Cook, W. J. (2012). Two proofs of the existence and uniqueness of the partial fraction decomposition. In *International Mathematical Forum*, volume 7, pages 1517–1535.

Browning, M. and Carro, J. M. (2014). Dynamic binary outcome models with maximal heterogeneity. *Journal of Econometrics*, 178(2):805–823.

Card, D. and Hyslop, D. R. (2005). Estimating the effects of a time-limited earnings subsidy for welfare-leavers. *Econometrica*, 73(6):1723–1770.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The review of economic studies*, 47(1):225–238.

Chamberlain, G. (1985). *Heterogeneity, omitted variable bias, and duration dependence*, page 3–38. Econometric Society Monographs. Cambridge University Press.

Chay, K. Y., Hoynes, H. W., and Hyslop, D. (1999). A non-experimental analysis of true state dependence in monthly welfare participation sequences. In *American Statistical Association*, pages 9–17.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Dano, K., Honoré, B. E., and Weidner, M. (2025). Dynamic panel multinomial logit models.

Davezies, L., D'Haultfœuille, X., and Mugnier, M. (2023). Fixed-effects binary choice models with three or more periods. *Quantitative Economics*, 14(3):1105–1132.

Deza, M. (2015). Is there a stepping stone effect in drug use? separating state dependence from unobserved heterogeneity within and between illicit drugs. *Journal of Econometrics*, 184(1):193–207.

Dobronyi, C., Gu, J., et al. (2021). Identification of dynamic panel logit models with fixed effects. *arXiv preprint arXiv:2104.04590*.

Graham, B. S. (2013). Comment on "social networks and the identification of peer effects" by paul goldsmith-pinkham and guido w. imbens. *Journal of Business and Economic Statistics*, 31(3):266–270.

Graham, B. S. (2016). Homophily and transitivity in dynamic network formation. Technical report, National Bureau of Economic Research.

Gu, J., Hahn, J., and Kim, K. I. (2023). The information bound of a dynamic panel logit model with fixed effects—corrigendum. *Econometric Theory*, 39(1):219–219.

Hahn, J. (1994). The efficiency bound of the mixed proportional hazard model. *The Review of Economic Studies*, 61(4):607–629.

Hahn, J. (1997). A note on the efficient semiparametric estimation of some exponential panel models. *Econometric Theory*, 13(4):583–588.

Hahn, J. (2001). The information bound of a dynamic panel logit model with fixed effects. *Econometric Theory*, 17(5):913–932.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Honoré, B. E., Hu, L., Kyriazidou, E., and Weidner, M. (2022). Simultaneity in binary outcome models with an application to employment for couples. *arXiv preprint arXiv:2207.07343*.

Honoré, B. E. and Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68(4):839–874.

Honoré, B. E. and Kyriazidou, E. (2019). Panel vector autoregressions with binary data. In *Panel Data Econometrics*, pages 197–223. Elsevier.

Honoré, B. E., Muris, C., and Weidner, M. (2021). Dynamic ordered panel logit models. *arXiv preprint arXiv:2107.03253*.

Honoré, B. E. and Weidner, M. (2020). Moment conditions for dynamic panel logit models with fixed effects. *arXiv preprint arXiv:2005.05942*.

Imbens, G. W. (1997). One-step estimators for over-identified generalized method of moments models. *The Review of Economic Studies*, 64(3):359–383.

Kitazawa, Y. (2022). Transformations and moment conditions for dynamic fixed effects logit models. *Journal of Econometrics*, 229(2):350 – 362.

Kitazawa, Y. et al. (2013). Exploration of dynamic fixed effects logit models from a traditional angle. Technical report.

Kitazawa, Y. et al. (2016). Root-n consistent estimations of time dummies for the dynamic fixed effects logit models: Monte carlo illustrations. Technical report.

Kruiniger, H. (2020). Further results on the estimation of dynamic panel logit models with fixed effects. *arXiv preprint arXiv:2010.03382*.

Lang, S. (2012). *Algebra*, volume 211. Springer Science & Business Media.

Magnac, T. (2000). Subsidised training and youth employment: distinguishing unobserved heterogeneity from state dependence in labour market histories. *The economic journal*, 110(466):805–837.

Muris, C., Raposo, P., and Vandoros, S. (2020). A dynamic ordered logit model with fixed effects. *arXiv preprint arXiv:2008.05517*.

Narendranthan, W., Nickell, S., and Metcalf, D. (1985). An investigation into the incidence and dynamic structure of sickness and unemployment in britain, 1965–75. *Journal of the Royal Statistical Society: Series A (General)*, 148(3):254–267.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of applied econometrics*, 5(2):99–135.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32.

Pakes, A., Porter, J. R., Shepard, M., and Calder-Wang, S. (2021). Unobserved heterogeneity, state dependence, and health plan choices. Technical report, National Bureau of Economic Research.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *the Annals of Statistics*, 22(1):300–325.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, 90(1):77–97.

Appendix

A Partial Fraction Decomposition

Lemma 6. *For any reals $u_1, u_2, \ldots, u_K$, $v_1, v_2, \ldots, v_K$ and $a_1, a_2, \ldots, a_K$, $K \geq 1$ we have*

$$\frac{1}{1 + \sum_{k=1}^{K} e^{v_k + a_k}} + \sum_{k=1}^{K} (1 - e^{u_k - v_k}) \frac{e^{v_k + a_k}}{\left(1 + \sum_{k=1}^{K} e^{v_k + a_k}\right)\left(1 + \sum_{k=1}^{K} e^{u_k + a_k}\right)} = \frac{1}{1 + \sum_{k=1}^{K} e^{u_k + a_k}}$$

and

$$\frac{e^{v_j + a_j}}{1 + \sum_{k=1}^{K} e^{v_k + a_k}} + (1 - e^{-u_j + v_j}) \frac{e^{u_j + a_j}}{\left(1 + \sum_{k=1}^{K} e^{v_k + a_k}\right)\left(1 + \sum_{k=1}^{K} e^{u_k + a_k}\right)} +$$

$$\sum_{\substack{k=1 \\ k \neq j}}^{K} (1 - e^{(u_k - u_j) - (v_k - v_j)}) \frac{e^{v_k + a_k + u_j + a_j}}{\left(1 + \sum_{k=1}^{K} e^{v_k + a_k}\right)\left(1 + \sum_{k=1}^{K} e^{u_k + a_k}\right)} = \frac{e^{u_j + a_j}}{1 + \sum_{k=1}^{K} e^{u_k + a_k}}$$

Proof. Verification of these identities is straightforward and thus left to the reader. □

Lemma 7. *Fix $M \geq 2$, let $\mathcal{Y} = \{0,1\}^M$. Then, for any $k \in \mathcal{Y}$ and any reals $u_1, u_2, \ldots, u_M$, $v_1, v_2, \ldots, v_M$ and $a_1, a_2, \ldots, a_M$, we have*

$$\prod_{m=1}^{M} \frac{e^{k_m(v_m+a_m)}}{1+e^{v_m+a_m}} + \sum_{l \in \mathcal{Y}\setminus\{k\}} \left[1 - e^{\sum_{j=1}^{M}(l_j-k_j)(u_j-v_j)}\right] \prod_{m=1}^{M} \frac{e^{k_m(u_m+a_m)}}{1+e^{u_m+a_m}} \frac{e^{l_m(v_m+a_m)}}{1+e^{v_m+a_m}} = \prod_{m=1}^{M} \frac{e^{k_m(u_m+a_m)}}{1+e^{u_m+a_m}}$$

Proof. Let

$$LHS = \prod_{m=1}^{M} \frac{e^{k_m(v_m+a_m)}}{1+e^{v_m+a_m}} + \sum_{l \in \mathcal{Y}\setminus\{k\}} \left[1 - e^{\sum_{j=1}^{M}(l_j-k_j)(u_j-v_j)}\right] \prod_{m=1}^{M} \frac{e^{k_m(u_m+a_m)}}{1+e^{u_m+a_m}} \frac{e^{l_m(v_m+a_m)}}{1+e^{v_m+a_m}}$$

and let Num denote the numerator of LHS. We have $Num = Num_1 + Num_2$ with

$$Num_1 = \prod_{m=1}^{M} e^{k_m(v_m+a_m)}(1+e^{u_m+a_m})$$

$$\begin{aligned}
Num_2 &= \sum_{l \in \mathcal{Y}\setminus\{k\}} \left[1 - e^{\sum_{j=1}^{M}(l_j-k_j)(u_j-v_j)}\right] \prod_{m=1}^{M} e^{k_m(u_m+a_m)+l_m(v_m+a_m)} \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)} \sum_{l \in \mathcal{Y}\setminus\{k\}} \prod_{m=1}^{M} e^{l_m(v_m+a_m)} - \sum_{l \in \mathcal{Y}\setminus\{k\}} e^{\sum_{j=1}^{M} l_j(u_j+a_j)+k_j(v_j+a_j)} \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)} \sum_{l \in \mathcal{Y}\setminus\{k\}} \prod_{m=1}^{M} e^{l_m(v_m+a_m)} - \prod_{m=1}^{M} e^{k_m(v_m+a_m)} \sum_{l \in \mathcal{Y}\setminus\{k\}} \prod_{m=1}^{M} e^{l_m(u_m+a_m)}
\end{aligned}$$

Since $\sum_{l \in \mathcal{Y}} \prod_{m=1}^{M} e^{l_m(v_m+a_m)} = \prod_{m=1}^{M}(1+e^{v_m+a_m})$, $\sum_{l \in \mathcal{Y}} \prod_{m=1}^{M} e^{l_m(u_m+a_m)} = \prod_{m=1}^{M}(1+e^{u_m+a_m})$ we get

$$\begin{aligned}
Num_2 = \prod_{m=1}^{M} e^{k_m(u_m+a_m)} &\left(\prod_{m=1}^{M}(1+e^{v_m+a_m}) - \prod_{m=1}^{M} e^{k_m(v_m+a_m)}\right) \\
&- \prod_{m=1}^{M} e^{k_m(v_m+a_m)} \left(\prod_{m=1}^{M}(1+e^{u_m+a_m}) - \prod_{m=1}^{M} e^{k_m(u_m+a_m)}\right) \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)}(1+e^{v_m+a_m}) - \prod_{m=1}^{M} e^{k_m(v_m+a_m)}(1+e^{u_m+a_m}) \\
&= \prod_{m=1}^{M} e^{k_m(u_m+a_m)}(1+e^{v_m+a_m}) - Num_1
\end{aligned}$$

It follows that $Num = \prod_{m=1}^M e^{k_m(u_m+a_m)}(1+e^{v_m+a_m})$ and consequently

$$LHS = \frac{\prod_{m=1}^M e^{k_m(u_m+a_m)}(1+e^{v_m+a_m})}{\prod_{m=1}^M (1+e^{u_m+a_m})(1+e^{v_m+a_m})} = \prod_{m=1}^M \frac{e^{k_m(u_m+a_m)}}{1+e^{u_m+a_m}}$$

□

B Proofs of key results in the main text

Proofs of Theorem 1 and Theorem 3. We focus on establishing Theorem 3 but highlight where the arguments for the AR(1) would differ at each important step of the proof. Fix a history $y \in \mathcal{Y}^T$ and consider the corresponding basis element $\mathbb{1}\{. = y\}$ of $\mathbb{R}^{\mathcal{Y}^T}$. We have: $\mathcal{E}^{(p)}_{y^0,x,T}\left[\mathbb{1}\{. = y\}\right] = P(Y_i = y|Y_i^0 = y^0, X_i = x, A_i = .)$ where by definition, for all $a \in \mathbb{R}$, $P(Y_i = y|Y_i^0 = y^0, X_i = x, A_i = a) = \frac{N^y(e^a)}{D^y(e^a)}$ with $N^y(e^a) = \prod_{t=1}^T e^{y_t\left(\sum_{r=1}^p \gamma_{0r} y_{t-r} + x_t'\beta_0 + a\right)}$ and $D^y(e^a) = \prod_{t=1}^T \left(1 + e^{\sum_{r=1}^p \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}\right)$. Notice that $N^y(e^a)$ and $D^y(e^a)$ are just polynomials of e^a - with dependence on y^0, x, T suppressed for conciseness - and that we always have $\deg\left(N^y(e^a)\right) \leq \deg\left(D^y(e^a)\right)$ with strict inequality unless $y = 1_T$. Moreover, since by assumption for any $t, s \in \{1, \ldots, T-1\}$ and $y, \tilde{y} \in \mathcal{Y}^p$, $\gamma_0' y + x_t'\beta_0 \neq \gamma_0'\tilde{y} + x_s'\beta_0$ if $t \neq s$ or $y \neq \tilde{y}$, $D^y(e^a)$ is a product of distinct irreducible polynomials in e^a. Thus, by standard results on partial fraction decompositions (e.g Bradley and Cook (2012), Theorem 5.2 in Lang (2012)), there exists a unique set of coefficients $(\lambda_0^y, \lambda_1^y, \ldots, \lambda_T^y) \in \mathbb{R}^{T+1}$ independent of the fixed effect such that:

$$\begin{aligned} P(Y_i = y|Y_i^0 = y^0, X_i = x, A_i = a) &= \lambda_0^y + \sum_{t=1}^T \lambda_t^y \frac{1}{1 + e^{\sum_{r=1}^p \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}} \\ &= \lambda_0^y + T_0(a) + T_1(a) + T_2(a) \end{aligned}$$

where $T_0(a) = \lambda_1^y \frac{1}{1+e^{\sum_{r=1}^p \gamma_{0r} y_{1-r} + x_1'\beta_0 + a}}$, $T_1(a) = \sum_{t=2}^p \lambda_t^y \frac{1}{1+e^{\sum_{r=1}^p \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}}$, and finally $T_2(a) = \sum_{t=p+1}^T \lambda_t^y \frac{1}{1+e^{\sum_{r=1}^p \gamma_{0r} y_{t-r} + x_t'\beta_0 + a}}$ with $\lambda_0^y = 0$ unless $y = 1_T$. In effect, this is merely the result of repeatedly applying Lemma 6 for the univariate fixed effect case $K = 1$. This decomposition breaks down the conditional probability $P(Y_i = y|Y_i^0 = y^0, X_i = x, A_i = a)$ into components that depend on the initial condition, namely $T_0(a), T_1(a)$, and components that do not, i.e $T_2(a)$. Notice that $T_1(a)$

would not appear in the AR(1) case. Starting with the first group, we can write:

$$
\begin{aligned}
T_0(a) &= \lambda_1^y \mathbb{1}\{y_0=1\} + \lambda_1^y \mathbb{1}\{y_0=0\}\pi_0^{y_0|y^0}(x,a) - \lambda_1^y \mathbb{1}\{y_0=1\}\pi_0^{y_0|y^0}(x,a) \\
T_1(a) &= \sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_2^{t-2}\in\mathcal{Y}^{t-2}} \mathbb{1}\{y_{t-1}=1, y_{t-2}=\tilde{y}_2,\dots,y_1=\tilde{y}_{t-1}\} \\
&\quad + \sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_2^{t-2}\in\mathcal{Y}^{t-2}} \mathbb{1}\{y_{t-1}=0, y_{t-2}=\tilde{y}_2,\dots,y_1=\tilde{y}_{t-1}\}\pi_{t-1}^{0|0,\tilde{y}_2^{t-1},y_0,\dots,y_{-(p-t)}}(a,x) \\
&\quad - \sum_{t=2}^{p} \lambda_t^y \sum_{\tilde{y}_2^{t-2}\in\mathcal{Y}^{t-2}} \mathbb{1}\{y_{t-1}=1, y_{t-2}=\tilde{y}_2,\dots,y_1=\tilde{y}_{t-1}\}\pi_{t-1}^{1|1,\tilde{y}_2^{t-1},y_0,\dots,y_{-(p-t)}}(a,x)
\end{aligned}
$$

Likewise, for the second group,

$$
\begin{aligned}
T_2(a) &= \sum_{t=p+1}^{T} \lambda_t^y \sum_{\tilde{y}_2^{p-1}\in\mathcal{Y}^{p-1}} \mathbb{1}\{y_{t-1}=1, y_{t-2}=y_2,\dots,y_{t-p}=\tilde{y}_p\} \\
&\quad + \sum_{t=p+1}^{T} \lambda_t^y \sum_{\tilde{y}_2^{p}\in\mathcal{Y}^{p-1}} \mathbb{1}\{y_{t-1}=0, y_{t-2}=y_2,\dots,y_{t-p}=\tilde{y}_p\}\pi_{t-1}^{0|0,\tilde{y}_2^{p}}(a,x) \\
&\quad - \sum_{t=p+1}^{T} \lambda_t^y \sum_{\tilde{y}_2^{p-1}\in\mathcal{Y}^{p-1}} \mathbb{1}\{y_{t-1}=1, y_{t-2}=y_2,\dots,y_{t-p}=\tilde{y}_p\}\pi_{t-1}^{1|1,\tilde{y}_2^{p}}(a,x)
\end{aligned}
$$

The above decompositions for each term make it clear that $\operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right) \subseteq \operatorname{span}(\mathcal{F}^{(p)}_{y^0,x,T})$ where

$$
\mathcal{F}^{(p)}_{y^0,x,T} = \left\{1, \pi_0^{y_0|y^0}(.,x), \left\{\left(\pi_{t-1}^{y_1|y_1^{t-1},y_0,\dots,y_{-(p-t)}}(.,x))\right)_{y_1^{t-1}\in\mathcal{Y}^{t-1}}\right\}_{t=2}^{p}, \left\{\left(\pi_{t-1}^{y_1|y_1^p}(.,x)\right)_{y_1^p\in\mathcal{Y}^p}\right\}_{t=p+1}^{T}\right\}
$$

We now argue that the converse holds: $\operatorname{span}(\mathcal{F}^{(p)}_{y^0,x,T}) \subseteq \operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)$.

- First, $\pi_0^{y_0|y^0}(.,x) \in \operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)$ since

$$
\mathbb{E}[(1-y_0)(1-Y_{i1}) + y_0 Y_{i1} | Y_i^0 = y^0, X_i = x, A_i = a] = \pi_0^{y_0|y^0}(a,x)
$$

- Second, $\left\{\left(\pi_{t-1}^{y_1|y_1^p}(.,x)\right)_{y_1^p\in\mathcal{Y}^p}\right\}_{t=p+1}^{T} \in \operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)$ by Theorem 4. For the AR(1) model, one would appeal to Lemma 2.

- Finally, one can easily adapt the reasoning employed to prove Theorem 4 to show that $\left\{\left(\pi_{t-1}^{y_1|y_1^{t-1},y_0,\ldots,y_{-(p-t)}}(.,x)\right)_{y_1^{t-1}\in\mathcal{Y}^{t-1}}\right\}_{t=2}^{p} \in \operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)$. In proving Theorem 4, we already established that: $\left(\pi_1^{y_1|y_1,y_0,\ldots,y_{-(p-2)}}(.,x)\right)_{y_1\in\mathcal{Y}^{t-1}} \in \operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)$. Now, by inspecting the induction argument of Theorem 4, it is easily seen that the result that for $T \geq p+1$ and $t \in \{p,\ldots,T-1\}$

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{k+1}}(Y_{it+1},Y_{it},Y_{it-(p+k)}^{t-1},X_i)|Y_i^0,Y_{i1}^{t-(k+1)},X_i,A_i\right] = \pi_t^{y_1|y_1^{k+1},Y_{it-(k+1)},\ldots,Y_{it-(p-1)}}(A_i,X_i)$$

for $k = 0,\ldots,p-2$ can be generalized. It actually holds for $t = k+1$ when $k = 0,\ldots,p-2$, yielding

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^t}(Y_{it+1},Y_{it},Y_{i1-p}^{t-1},X_i)|Y_i^0,X_i,A_i\right] = \pi_t^{y_1|y_1^t,Y_{i0},\ldots,Y_{it-(p-1)}}(A_i,X_i)$$

which is the desired result. These terms are not present in the AR(1) case which simplifies the argument.

Thus, we have shown that $\operatorname{span}(\mathcal{F}^{(p)}_{y^0,x,T}) = \operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)$. Next, a key observation is that the elements of $\mathcal{F}^{(p)}_{y^0,x,T}$ are linearly independent, and hence form a basis for $\operatorname{Im}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right)$. This follows from noting that the transition probabilities in $\mathcal{F}^{(p)}_{y^0,x,T}$ are rational fractions in e^a of the form $\frac{1}{1+u_k e^a}$ or $\frac{u_k e^a}{1+u_k e^a}$, for some $u_k \in \mathbb{R}_{++}$ that are all distinct (by assumption), and that a collection of such rational fractions is always linearly independent. See Lemma 8 in the Supplemental Appendix for a simple proof. Now, since $\mathcal{E}^{(p)}_{y^0,x,T}$ is a linear mapping, the *rank nullity theorem* entails: $\dim\left(\ker(\mathcal{E}^{(p)}_{y^0,x,T})\right) = \dim\left(\mathbb{R}^{\{0,1\}^T}\right) - \operatorname{rank}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right) = 2^T - |\mathcal{F}^{(p)}_{y^0,x,T}|$. We have the following implications:

1. If $T \leq p$, $|\mathcal{F}^{(p)}_{y^0,x,T}| = 1+1+\sum_{t=2}^{T} 2^{t-1} = 2+\sum_{t=1}^{T-1} 2^t = 2+2\frac{1-2^{T-1}}{1-2} = 2^T$. Hence, $\operatorname{rank}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right) = 2^T$ and $\dim\left(\ker(\mathcal{E}^{(p)}_{y^0,x,T})\right) = 2^T - 2^T = 0$

2. If $T = p+1$, $|\mathcal{F}_{y^0,p,T}| = 1+1+\sum_{t=2}^{p} 2^{t-1} + 2^p = 2\times 2^p = 2^{p+1}$. Then, $\operatorname{rank}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right) = 2^T$ and $\dim\left(\ker(\mathcal{E}^{(p)}_{y^0,x,T})\right) = 2^T - 2^{p+1} = 0$

3. If $T \geq p+2$, $|\mathcal{F}_{y^0,p,T}| = 1+1+\sum_{t=2}^{p} 2^{t-1}+2^p(T-p) = 2^p+2^p(T-p) = (T-p+1)2^p$. It follows that $\operatorname{rank}\left(\mathcal{E}^{(p)}_{y^0,x,T}\right) = (T-p+1)2^p$ and $\dim\left(\ker(\mathcal{E}^{(p)}_{y^0,x,T})\right) = 2^T - (T-p+1)2^p$

Proofs of Lemma 1 and Lemma 2. Without loss of generality, we will consider the case with covariates. The discussion in Section 3.2.1 implies the functional form $\phi_\theta^{k|k}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = \mathbb{1}\{Y_{it} = k\}\phi_\theta^{k|k}(Y_{it+1}, k, Y_{it-1}, X_i)$ for $k \in \mathcal{Y}$. Therefore, $\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)$ is null when $Y_{it} \neq 0$ implying

$$\mathbb{E}\left[\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \frac{1}{1 + e^{\gamma_0 Y_{it-1} + X_{it}'\beta_0 + A_i}} \times \left(\frac{e^{X_{it+1}'\beta_0 + A_i}}{1 + e^{X_{it+1}'\beta_0 + A_i}}\phi_\theta^{0|0}(1, 0, Y_{it-1}, X_i) + \frac{1}{1 + e^{X_{it+1}'\beta_0 + A_i}}\phi_\theta^{0|0}(0, 0, Y_{it-1}, X_i)\right)$$

Thus, to get the transition probability $\pi_t^{0|0}(A_i, X_i) = \frac{1}{1+e^{X_{it+1}'\beta_0 + A_i}}$ at $\theta = \theta_0$, it must be that $\phi_\theta^{0|0}(1, 0, Y_{it-1}, X_i) = e^{\gamma Y_{it-1} + (X_{it} - X_{it+1})'\beta}$, $\phi_\theta^{0|0}(0, 0, Y_{it-1}, X_i) = 1$, and that $\forall k \in \mathcal{Y}$ $\phi_\theta^{0|0}(k, 1, Y_{it-1}, X_i) = 0$. That is: $\phi_\theta^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = (1 - Y_{it})e^{Y_{it+1}\left(\gamma Y_{it-1} - \Delta X_{it+1}'\beta\right)}$.

Likewise, $\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)$ is null when $Y_{it} \neq 1$ implying

$$\mathbb{E}\left[\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right] = \frac{e^{\gamma_0 Y_{it-1} + X_{it}'\beta_0 + A_i}}{1 + e^{\gamma_0 Y_{it-1} + X_{it}'\beta_0 + A_i}} \times \left(\frac{e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}}{1 + e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}}\phi_\theta^{1|1}(1, 1, Y_{it-1}, X_i) + \frac{1}{1 + e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}}\phi_\theta^{1|1}(0, 1, Y_{it-1}, X_i)\right)$$

Hence, to get $\pi_t^{1|1}(A_i, X_i) = \frac{e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}}{1+e^{\gamma_0 + X_{it+1}'\beta_0 + A_i}}$ at $\theta = \theta_0$, we must set: $\phi_\theta^{1|1}(1, 1, Y_{it-1}, X_i) = 1$, $\phi_\theta^{1|1}(0, 1, Y_{it-1}, X_i) = e^{\gamma(1 - Y_{it-1}) + (X_{it+1} - X_{it})'\beta}$ and $\phi_\theta^{1|1}(k, 0, Y_{it-1}, X_i) = 0, \quad \forall k \in \mathcal{Y}$. In compact form this is: $\phi_\theta^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i) = Y_{it}e^{(1 - Y_{it+1})\left(\gamma(1 - Y_{it-1}) + \beta\Delta X_{it+1}\right)}$

Proof of Lemma 3. By construction for $T \geq 3$, and t, s such that $T - 1 \geq t > s \geq 1$:

$$\begin{aligned}
&\mathbb{E}\left[\zeta_{\theta_0}^{0|0}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \mathbb{E}\left[(1 - Y_{is}) + \omega_{t,s}^{0|0}(\theta_0)Y_{is}\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{1}{1 + e^{\mu_s(\theta_0) + A_i}} + \omega_{t,s}^{0|0}(\theta_0)\mathbb{E}\left[Y_{is}\mathbb{E}\left[\phi_{\theta_0}^{0|0}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{1}{1 + e^{\mu_s(\theta_0) + A_i}} + \omega_{t,s}^{0|0}(\theta_0)\mathbb{E}\left[Y_{is}|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right]\frac{1}{1 + e^{\kappa_t^{0|0}(\theta_0) + A_i}} \\
&= \frac{1}{1 + e^{\mu_s(\theta_0) + A_i}} + (1 - e^{\kappa_t^{0|0}(\theta_0) - \mu_s(\theta_0)})\frac{e^{\mu_s(\theta_0) + A_i}}{(1 + e^{\mu_s(\theta_0) + A_i})(1 + e^{\kappa_t^{0|0}(\theta_0) + A_i})} \\
&= \frac{1}{1 + e^{\kappa_t^{0|0}(\theta_0) + A_i}} \\
&= \pi_t^{0|0}(A_i, X_i)
\end{aligned}$$

The second equality follows from the measurability of the weight $\omega_{t,s}^{0|0}(\theta_0)$ with respect to the conditioning set. The third equality follows from the law of iterated expectations and Lemma 2. The penultimate equality uses the first identity in Lemma 6 (for $K = 1$) . Similarly,

$$
\begin{aligned}
&\mathbb{E}\left[\zeta_{\theta_0}^{1|1}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \mathbb{E}\left[Y_{is} + \omega_{t,s}^{1|1}(\theta_0)(1-Y_{is})\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{e^{\mu_s(\theta_0)+A_i}}{1+e^{\mu_s(\theta_0)+A_i}} + \omega_{t,s}^{1|1}(\theta_0)\mathbb{E}\left[(1-Y_{is})\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1}, Y_{it}, Y_{it-1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \frac{e^{\mu_s(\theta_0)+A_i}}{1+e^{\mu_s(\theta_0)+A_i}} + \omega_{t,s}^{1|1}(\theta_0)\mathbb{E}\left[(1-Y_{is})|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right]\frac{e^{\kappa_t^{1|1}(\theta_0)+A_i}}{1+e^{\kappa_t^{1|1}(\theta_0)+A_i}} \\
&= \frac{e^{\mu_s(\theta_0)+A_i}}{1+e^{\mu_s(\theta_0)+A_i}} + \left(1-e^{-(\kappa_t^{1|1}(\theta_0)-\mu_s(\theta_0))}\right)\frac{e^{\kappa_t^{1|1}(\theta_0)+A_i}}{(1+e^{\mu_s(\theta_0)+A_i})(1+e^{\kappa_t^{1|1}(\theta_0)+A_i})} \\
&= \frac{e^{\kappa_t^{1|1}(\theta_0)+A_i}}{1+e^{\kappa_t^{1|1}(\theta_0)+A_i}} \\
&= \pi_t^{1|1}(A_i, X_i)
\end{aligned}
$$

The second equality follows from the measurability of the weight $\omega_{t,s}^{1|1}(\theta_0)$ with respect to the conditioning set. The third equality follows from the law of iterated expectations and Lemma 2. The penultimate equality uses the second identity in Lemma 6 (for $K = 1$). Showing $\mathbb{E}\left[\zeta_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i)|Y_{i0}, Y_{i1}^{s_J-1}, X_i, A_i\right] = \pi_t^{k|k}(A_i, X_i)$ is analogous.

Proof of Proposition 1. For any t, s verifying $T-1 \geq t > s \geq 1$ and $k \in \mathcal{Y}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\psi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s+1}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, X_i) - \zeta_{\theta}^{k|k}(Y_{it-1}^{t+1}, Y_{is-1}^{s}, X_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\phi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, X_i)|Y_{i0}, Y_{i1}^{t-1}, X_i, A_i\right]|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] - \pi_t^{k|k}(A_i, X_i) \\
&= \mathbb{E}\left[\pi_t^{k|k}(A_i)|Y_{i0}, Y_{i1}^{s-1}, X_i, A_i\right] - \pi_t^{k|k}(A_i) \\
&= \pi_t^{k|k}(A_i) - \pi_t^{k|k}(A_i) \\
&= 0
\end{aligned}
$$

The second and third equalities follow from the law of iterated expectations, Lemma

3 and Lemma 2. Showing $\mathbb{E}\left[\psi_{\theta_0}^{k|k}(Y_{it-1}^{t+1}, Y_{is_1-1}^{s_1}, \ldots, Y_{is_J-1}^{s_J}, X_i)|Y_{i0}, Y_{i1}^{s_J-1}, X_i, A_i\right] = 0$ is analogous.

Proof of Proposition 2. In what follows, we drop the cross-sectional subscript i to economize on space. The proof is based on an application of Theorem 3.2 in Newey (1990). First, in paragraphs I)-II), we verify that the model is mean-square differentiable and characterize the nonparametric tangent set $\mathcal{T}$. Second, in paragraphs III)-IV), we characterize its orthogonal complement $\mathcal{T}^\perp$ and verify that the efficient score - the projection of the score onto $\mathcal{T}^\perp$ - coincides with the efficient moment function for $\mathbb{E}\left[\psi_{\theta_0}(Y_0, Y, X)|Y_0, X\right] = 0$, namely $\psi_{\theta_0}^{eff}(Y_0, Y, X) = -D(Y_0, X)'\Sigma(Y_0, X)^{-1}\psi_{\theta_0}(Y_0, Y, X)$ (see e.g Newey and McFadden (1994))

I) Preliminary calculations

The parametric component of the AR(1) model writes $f(Y|Y_0, X, A; \theta) = \prod_{t=1}^{T} \frac{e^{Y_t(\gamma Y_{t-1} + X_t'\beta + A)}}{\left(1+e^{\gamma Y_{t-1} + X_t'\beta + A}\right)}$. This implies

$$\begin{aligned}
\ln f(Y|Y_0, X, A; \theta) = \sum_{t=1}^{T} Y_t(\gamma Y_{t-1} + X_t'\beta + A) - \sum_{t=1}^{T} Y_{t-1} \ln\left(1 + e^{\gamma + X_t'\beta + A}\right) \\
- \sum_{t=1}^{T} (1 - Y_{t-1}) \ln\left(1 + e^{X_t'\beta + A}\right)
\end{aligned}$$

which is continously differentiable in θ. Hence

$$\begin{aligned}
\frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \gamma} &= \sum_{t=1}^{T} Y_{t-1}\left(Y_t - \frac{e^{\gamma + X_t'\beta + A}}{1 + e^{\gamma + X_t'\beta + A}}\right) \\
\frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \beta} &= \sum_{t=1}^{T} X_t\left(Y_t - Y_{t-1}\frac{e^{\gamma + X_t'\beta + A}}{1 + e^{\gamma + X_t'\beta + A}} - (1 - Y_{t-1})\frac{e^{X_t'\beta + A}}{1 + e^{X_t'\beta + A}}\right)
\end{aligned}$$

and because $\mathcal{Y} = \{0, 1\}$, we have

$$\begin{aligned}
\left|\frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \gamma}\right| &\leq T \\
\left|\frac{\partial \ln f(Y|Y_0, X, A; \theta)}{\partial \beta}\right| &\leq \sum_{t=1}^{T} |X_t|
\end{aligned} \tag{4}$$

II) Mean-square differentiability and nonparametric tangent set

Consider a parametric likelihood for $(Y, A)|Y_0, X$

$$f(Y, A|Y_0, X; \theta, \eta) = f(Y|Y_0, X, A; \theta)q(A|Y_0, X; \eta)$$

where $q(.|Y_0, X; \eta)$ is a density for the heterogeneity such that: a) at $\eta = 0$, $q(.|Y_0, X; 0) = q(.|Y_0, X)$ and b) $q(.|Y_0, X; \eta)^{1/2}$ is mean-square differentiable at $\eta = 0$ with derivative equal to $\frac{1}{2}q(.|Y_0, X)K(.|Y_0, X)$. We will prove that $f(Y, A|Y_0, X; \theta, \eta)^{1/2}$ is mean-square differentiable at $(\theta_0, 0)$, meaning $\mathbb{E}[\Upsilon] = o(\|h^2\| + \eta^2)$ where

$$\begin{aligned}\Upsilon &= f(Y, A|Y_0, X; \theta_0 + h, \eta)^{1/2} - f(Y, A|Y_0, X; \theta_0, 0)^{1/2} \\ &\quad - \frac{1}{2}f(Y, A|Y_0, X; \theta_0, 0)^{1/2}\left(h'\frac{\partial \ln f(Y|Y_0, X, A; \theta_0)}{\partial \theta} + \eta K(A|Y_0, X)\right)\end{aligned}$$

Similarly to Lemma A-2 in Hahn (1994), we decompose Υ in three terms

$$\begin{aligned}\Upsilon &= \Upsilon_1 + \Upsilon_2 + \Upsilon_3 \\ \Upsilon_1 &= \left(q(A|Y_0, X; \eta)^{1/2} - q(A|Y_0, X)^{1/2} - \frac{\eta}{2}q(A|Y_0, X)^{1/2}K(A|Y_0, X)\right) f(Y|Y_0, X, A; \theta_0 + h)^{1/2} \\ \Upsilon_2 &= \left(f(Y|Y_0, X, A; \theta_0 + h)^{1/2} - f(Y|Y_0, X, A; \theta_0)^{1/2} - \frac{1}{2}f(Y|Y_0, X, A; \theta_0)^{1/2}h'\frac{\partial \ln f(Y|Y_0, X, A; \theta_0)}{\partial \theta}\right) \\ &\quad \times q(A|Y_0, X)^{1/2} \\ \Upsilon_3 &= \frac{\eta}{2}q(A|Y_0, X)^{1/2}K(A|Y_0, X)\left(f(Y|Y_0, X, A; \theta_0 + h)^{1/2} - f(Y|Y_0, X, A; \theta_0)^{1/2}\right)\end{aligned}$$

By Jensen's inequality, we have $\Upsilon^2 \leq 3\Upsilon_1^2 + 3\Upsilon_2^2 + 3\Upsilon_3^2$. By b), $\mathbb{E}[\Upsilon_1^2] = o(\eta^2) = o(\|h\|^2 + \eta^2)$. To show that $\mathbb{E}[\Upsilon_2^2] = o(\|h\|^2 + \eta^2)$, we verify that $f(.|Y_0, X, A; \theta_0)$ verifies the conditions of Lemma A-1 in Hahn (1994). The first condition is that $f(Y|Y_0, X, A; .)$ is continuously differentiable in θ which follows from paragraph I). The second condition is that $\mathbb{E}\left[\frac{\partial \ln f(Y|Y_0,X,A;.)}{\partial \theta}\frac{\partial \ln f(Y|Y_0,X,A;.)}{\partial \theta'}\right]$ is continuous in θ and finite at θ_0. This follows from Theorem 2 assumption i), inequalities (4) and the dominated convergence theorem. By Lemma A-1 in Hahn (1994), $f(.|Y_0, X, A; \theta)^{1/2}$ is mean square differentiable at θ_0 with derivative $\frac{1}{2}f(Y|Y_0, X, A; \theta_0)^{1/2}\frac{\partial \ln f(Y|Y_0,X,A;\theta_0)}{\partial \theta}$. This implies that: $\mathbb{E}\left[\Upsilon_2^2\right] = o(\|h\|^2) = o(\|h\|^2 + \delta^2)$. Last, $\mathbb{E}\left[\Upsilon_3^2\right] = o(\|h\|^2 + \delta^2)$ by the arguments on pages 624-625 of Hahn (1994). We conclude that $f(Y, A|Y_0, X; \theta, \eta)^{1/2}$ is mean-square differentiable at $(\theta_0, 0)$ with derivative:

$$\frac{1}{2}f(Y, A|Y_0, X; \theta_0, 0)^{1/2}\left(\frac{\partial \ln f(Y|Y_0, X, A; \theta_0)}{\partial \theta'}, K(A|Y_0, X)\right)'$$

From Bickel et al. (1993) Proposition A.5.5, $f(Y|Y_0, X; \theta, \eta)^{1/2}$ is mean-square differentiable at $(\theta_0, 0)$ with derivative

$$\frac{1}{2} f(Y|Y_0, X; \theta_0, 0)^{1/2} \left(\mathbb{E}\left[\frac{\partial \ln f(Y|Y_0, X, A; \theta_0)}{\partial \theta'} |Y_0, Y, X \right], \mathbb{E}\left[K(A|Y_0, X)|Y_0, Y, X \right] \right)'$$

This implies that the nonparametric tangent set is

$$\mathcal{T} = \left\{ \mathbb{E}[K(A, Y_0, X)|Y_0, Y, X] \text{ such that } \mathbb{E}[K(A, Y_0, X)|Y_0, X] = 0 \right\}$$

Having established mean-square differentiability of the model, noting that $\mathcal{T}$ is linear, and that by Theorem 2 assumption iii),

$$\mathbb{E}\left[\psi_{\theta_0}^{eff}(Y_0, Y, X) \psi_{\theta_0}^{eff}(Y_0, Y, X)' \right] = \mathbb{E}\left[D(Y_0, X)' \Sigma(Y_0, X)^{-1} D(Y_0, X)' \right]$$

is nonsingular, all that remains to check are: c) $\psi_{\theta_0}^{eff}(Y_0, Y, X) \in \mathcal{T}^{\perp}$ and d) $S^{\theta}(Y_0, Y, X) - \psi_{\theta_0}^{eff}(Y_0, Y, X) \in \mathcal{T}$ where $S^{\theta}(Y_0, Y, X) = \mathbb{E}\left[\frac{\partial \ln f(Y|Y_0, X, A; \theta_0)}{\partial \theta} |Y_0, Y, X \right]$. To this end, similarly to Hahn (1997), we shall first show that c) and d) hold conditional on a pair $(y_0, x) \in \mathcal{Y} \times \mathcal{X}^T$ for the initial condition and the regressors. In other words, we will prove next that $\psi_{\theta_0}^{eff}(y_0, Y, x)$ is the projection of the score onto the orthocomplement of the closed linear space

$$\mathcal{T}_{(y_0,x)} = \left\{ \mathbb{E}[K(A, y_0, x)|Y_0 = y_0, Y, X = x] \text{ such that } \mathbb{E}[K(A, y_0, x)|Y_0 = y_0, X = x] = 0 \right\}$$

in the Hilbert space of mean-zero square integrable random variables with inner product $< f_1, f_2 >= \mathbb{E}\left[f_1(y_0, Y, x)' f_2(y_0, Y, x)|Y_0 = y_0, X = x \right]$.

III) Verification of condition c) $\psi_{\theta_0}^{\mathbf{eff}}(\mathbf{y_0}, \mathbf{Y}, \mathbf{x}) \in \mathcal{T}_{(\mathbf{y_0},\mathbf{x})}^{\perp}$

We begin by characterizing the orthocomplement of the nonparametric tangent set $\mathcal{T}_{(y_0,x)}^{\perp}$. By definition, any $g(y_0, Y, x) \in \mathcal{T}_{(y_0,x)}^{\perp}$ is such that for any element of $\mathcal{T}_{(y_0,x)}$, $\mathbb{E}[K(A, y_0, x)|Y_0 = y_0, Y, X = x]$, we have $\mathbb{E}\left[g(y_0, Y, x)' \mathbb{E}[K(A, y_0, x)|Y_0 = y_0, Y, X = x]'|Y_0 = y_0, X = x \right] = 0$. By Lemma A.1 in Newey (1990), this is equivalent to

$$\begin{aligned} 0 &= \mathbb{E}\left[g(y_0, Y, x) \mathbb{E}[K(A, y_0, x)|Y_0 = y_0, Y, X = x]'|Y_0 = y_0, X = x \right] \\ &= \int \mathbb{E}\left[g(y_0, Y, x)|Y_0 = y_0, X = x, A = a \right] K(a, y_0, x)' q(a|y_0, x) da \end{aligned}$$

Since this equality must hold for any K verifying $\mathbb{E}[K(A, y_0, x)|Y_0 = y_0, X = x] = 0$, choosing

$$K(A,y_0,x)=\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x,A\right]-\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x\right]$$

yields $\mathbb{V}\left(\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x,A\right]|Y_0=y_0,X=x\right)=0$ so that $\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x,A\right]=c(y_0,x)$ q-a.s for some constant vector $c(y_0,x)$. In fact, it must be that $c(y_0,x)=\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x,A\right]=0$ q-a.s since by definition, $\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x\right]=0$. To see that this equality actually holds on the entire real line beyond $\mathcal{A}_q$ - the support of q - remark first that $\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x,A=.\right]$ is real analytic on $\mathcal{A}_q$ since logit probabilities are real analytic as ratios of exponential functions. Second, by Theorem 2 assumption ii), $\mathcal{A}_q$ has an accumulation point. Thus, the Identity Theorem (see e.g Proposition 7 in Argañaraz and Escanciano (2023)) implies that $\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x,A\right]=0,\ A\in\mathbb{R}$. Conversely, it is clear that any g function such that $\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x,A\right]=0$ for all $A\in\mathbb{R}$ will be an element of $\mathcal{T}^{\perp}_{(y_0,x)}$ since $\mathbb{E}[K(A,y_0,x)|Y_0=y_0,X=x]=0$. We conclude that

$$\mathcal{T}^{\perp}_{(y_0,x)}=\{g(y_0,Y,x)\,|\,\mathbb{E}\left[g(y_0,Y,x)|Y_0=y_0,X=x,A\right]=0,A\in\mathbb{R}\}$$

An important observation is that $\mathcal{T}^{\perp}_{(y_0,x)}=\ker(\mathcal{E}_{y_0,x,T})^{K_x+1}$, where we recall that the nullspace of the conditional expectation operator $\mathcal{E}_{y_0,x,T}$ is precisely the set of valid moment functions in the AR(1) model. By Theorem 1, this is a (2^T-2T)-dimensional vector space with an example of basis elements given in Proposition 1. This makes it clear that $\psi^{eff}_{\theta_0}(y_0,Y,x)\in\mathcal{T}^{\perp}_{(y_0,x)}$ since each of its components is a linear combination of the valid moment functions in Proposition 1.

IV) Verification of condition d) $\mathbf{S}^{\theta}(\mathbf{y_0},\mathbf{Y},\mathbf{x})-\psi^{\mathbf{eff}}_{\theta_0}(\mathbf{y_0},\mathbf{Y},\mathbf{x})\in\mathcal{T}_{(\mathbf{y_0},\mathbf{x})}$

Since $\mathcal{T}_{(y_0,x)}$ is a closed vector space of a Hilbert space, $\mathcal{T}_{(y_0,x)}=\left(\mathcal{T}^{\perp}_{(y_0,x)}\right)^{\perp}$. Thus, to check condition d) $S^{\theta}(y_0,Y,x)-\psi^{eff}_{\theta_0}(y_0,Y,x)\in\mathcal{T}_{(y_0,x)}$, we will verify that $\forall g\in\mathcal{T}^{\perp}_{(y_0,x)}$, $\mathbb{E}\left[\left(S^{\theta}(y_0,Y,x)-\psi^{eff}_{\theta_0}(y_0,Y,x)\right)g(y_0,Y,x)'|Y_0=y_0,X=x\right]=0$. Given our characterization of $\mathcal{T}^{\perp}_{(y_0,x)}$, it is sufficient to check that $\mathbb{E}\left[\left(S^{\theta}(Y_0,Y,X)-\psi^{eff}_{\theta_0}(y_0,Y,x)\right)\psi_{\theta_0}(y_0,Y,x)'|Y_0=y_0,X=x\right]=0$.

To this end, note that by the Generalized Information Equality (c.f equation (5.1) in

Newey and McFadden (1994)) we have

$$\mathbb{E}\left[\frac{\partial\psi_{\theta_0}(y_0,Y,x)}{\partial\theta'}|Y_0=y_0,X=x\right]=-\mathbb{E}\left[\psi_{\theta_0}(y_0,Y,x)S^\theta(y_0,Y,x)'|Y_0=y_0,X=x\right]$$

which implies

$$\begin{aligned}\psi^{eff}_{\theta_0}(y_0,Y,x)=&\ \mathbb{E}\left[\psi_{\theta_0}(y_0,Y,x)S^\theta(y_0,Y,x)'|Y_0=y_0,X=x\right]'\times\\&\mathbb{E}\left[\psi_{\theta_0}(y_0,Y,x)\psi_{\theta_0}(y_0,Y,x)'|Y_0=y_0,X=x\right]^{-1}\psi_{\theta_0}(y_0,Y,x)\\&=\mathbb{E}\left[S^\theta(y_0,Y,x)\psi_{\theta_0}(y_0,Y,x)'|Y_0=y_0,X=x\right]\times\\&\mathbb{E}\left[\psi_{\theta_0}(y_0,Y,x)\psi_{\theta_0}(y_0,Y,x)'|Y_0=y_0,X=x\right]^{-1}\psi_{\theta_0}(y_0,Y,x)\\&=\mathbb{E}^*\left[S^\theta(y_0,Y,x)|\psi_{\theta_0}(y_0,Y,x);Y_0=y_0,X=x\right]\end{aligned}$$

where $\mathbb{E}^*\left[Z_1|Z_2;W\right]$ denotes the (mean-squared error minimizing) linear predictor of Z_1 on Z_2 given W. Therefore, it immediately follows by properties of conditional linear predictors (e.g Wooldridge (1999), Lemma 4.1) that

$$\mathbb{E}\left[\left(S^\theta(y_0,Y,x)-\psi^{eff}_{\theta_0}(y_0,Y,x)\right)\psi_{\theta_0}(y_0,Y,x)'|Y_0=y_0,X=x\right]=0$$

We conclude that $\psi^{eff}_{\theta_0}(y_0,Y,x)$ is the projection of the score onto $\mathcal{T}^\perp_{(y_0,x)}$. It follows that $\psi^{eff}_{\theta_0}(Y_0,Y,X)$ is the projection of the score onto $\mathcal{T}^\perp$, i.e it is the efficient score.

Proof sketch of Theorem 4. In model (ARp), with $T\geq 2$ and $t\in\{1,\dots,T-1\}$, the moment functions

$$\begin{aligned}\phi^{0|0}_\theta(Y_{it+1},Y_{it},Y^{t-1}_{it-p},X_i)&=(1-Y_{it})e^{Y_{it+1}(\gamma_1Y_{it-1}-\sum_{l=2}^p\gamma_l\Delta Y_{it+1-l}-\Delta X'_{it+1}\beta)}\\\phi^{1|1}_\theta(Y_{it+1},Y_t,Y^{t-1}_{it-p},X_i)&=Y_{it}e^{(1-Y_{it+1})\left(\gamma_1(1-Y_{it-1})+\sum_{l=2}^p\gamma_l\Delta Y_{it+1-l}+\Delta X'_{it+1}\beta\right)}\end{aligned}$$

can be viewed as the counterpart of the AR(1) transition functions in Lemma 2 where one would treat lagged outcome variables Y_{it-r} for $r=2,\dots,p$ as additional strictly exogenous regressors. Leveraging this insight, it immediately follows from the proof of Lemma 2 that

$$\begin{aligned}\mathbb{E}\left[\phi^{0|0}_{\theta_0}(Y_{it+1},Y_{it},Y^{t-1}_{it-p},X_i)|Y^0_i,Y^{t-1}_{i1},X_i,A_i\right]&=\pi_t^{0|0,Y_{it-1},\dots,Y_{it-(p-1)}}(A_i,X_i)\\&=\frac{1}{1+e^{\sum_{l=2}^p\gamma_{0l}Y_{it+1-l}+X'_{it+1}\beta_0+A_i}}\end{aligned}$$

$$\mathbb{E}\left[\phi_{\theta_0}^{1|1}(Y_{it+1},Y_{it},Y_{it-p}^{t-1},X_i)|Y_i^0,Y_{i1}^{t-1},X_i,A_i\right]=\pi_t^{1|1,Y_{it-1},\dots,Y_{it-(p-1)}}(A_i,X_i)$$
$$=\frac{e^{\gamma_{01}+\sum_{l=2}^p\gamma_{0l}Y_{it+1-l}+X_{it+1}'\beta_0+A_i}}{1+e^{\gamma_{01}+\sum_{l=2}^p\gamma_{0l}Y_{it+1-l}+X_{it+1}'\beta_0+A_i}}$$

Now, for $T\geq p+1$ fix $t\in\{p,\dots,T-1\}$ and $y=(y_1,\dots,y_p)=y_1^p\in\{0,1\}^p$. One can show by finite induction the statement $\mathcal{P}(k)$:

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{k+1}}(Y_{it+1},Y_{it},Y_{it-(p+k)}^{t-1},X_i)|Y_i^0,Y_{i1}^{t-(k+1)},X_i,A_i\right]=\pi_t^{y_1|y_1^{k+1},Y_{it-(k+1)},\dots,Y_{it-(p-1)}}(A_i,X_i)$$

for $k=0,\dots,p-2$, $p\geq 2$. We give a brief proof sketch below.

Base step: $\mathcal{P}(0)$ is true by the above result which also deals with the edge case $p=2$. Thus, we can assume $p\geq 3$ in the remainder of the induction argument.

Induction step: Suppose $\mathcal{P}(k-1)$ is true for some $k\in\{1,\dots,p-2\}$, we show that $\mathcal{P}(k)$ is true. Using the law of iterated expectations, the induction hypothesis $\mathcal{P}(k-1)$ and the identities of Lemma 6, we have for $y_1=0,y_{k+1}=1$

$$\begin{aligned}
&\mathbb{E}\left[\phi_{\theta_0}^{0|0,y_2^k,1}(Y_{it+1},Y_{it},Y_{it-(p+k)}^{t-1},X_i)|Y_i^0,Y_{i1}^{t-(k+1)},X_i,A_i\right]\\
&=\mathbb{E}\left[(1-Y_{it-k})+w_t^{0|0,y_2^k,1}(\theta_0)\phi_{\theta_0}^{0|0,y_2^k}(Y_{it+1},Y_{it},Y_{it-(p+k-1)}^{t-1},X_i)Y_{it-k}|Y_i^0,Y_{i1}^{t-(k+1)},X_i,A_i\right]\\
&=\frac{1}{1+e^{u_{t-k}(\theta_0)+A_i}}\\
&+w_t^{0|0,y_2^k,1}(\theta_0)\mathbb{E}\left[\mathbb{E}\left[\phi_{\theta_0}^{0|0,y_2^k}(Y_{it+1},Y_{it},Y_{it-(p+k-1)}^{t-1},X_i)|Y_i^0,Y_{i1}^{t-k},X_i,A_i\right]Y_{it-k}|Y_i^0,Y_{i1}^{t-(k+1)},X_i,A_i\right]\\
&=\frac{1}{1+e^{u_{t-k}(\theta_0)+A_i}}w_t^{0|0,y_2^k,1}(\theta_0)\mathbb{E}\left[\pi_t^{0|0,y_2^k,Y_{it-k},\dots,Y_{it-(p-1)}}(A_i,X_i)Y_{it-k}|Y_i^0,Y_{i1}^{t-(k+1)},X_i,A_i\right]\\
&=\frac{1}{1+e^{u_{t-k}(\theta_0)+A_i}}+w_t^{0|0,y_2^k,1}(\theta_0)\mathbb{E}\left[\frac{1}{1+e^{\sum_{r=2}^k\gamma_{0r}y_r+\sum_{r=k+1}^p\gamma_{0r}Y_{it-(r-1)}+X_{it+1}'\beta_0+A_i}}Y_{it-k}|Y_i^0,Y_{i1}^{t-(k+1)},X_i,A_i\right]\\
&=\frac{1}{1+e^{u_{t-k}(\theta_0)+A_i}}+(1-e^{(k_t^{0|0,y_2^k,1}(\theta_0)-u_{t-k}(\theta_0))})\frac{1}{1+e^{k_t^{0|0,y_2^k,1}(\theta_0)+A_i}}\frac{e^{u_{t-k}(\theta_0)+A_i}}{1+e^{u_{t-k}(\theta_0)+A_i}}\\
&=\frac{1}{1+e^{k_t^{0|0,y_2^k,1}(\theta_0)+A_i}}\\
&=\pi_t^{0|0,y_2^k,1,Y_{it-(k+1)},\dots,Y_{it-(p-1)}}(A_i,X_i)
\end{aligned}$$

We leave out the derivations for the other three configurations: $y_1=0,y_{k+1}=0$, and $y_1=1,y_{k+1}=0$, and $y_1=1,y_{k+1}=1$ which follow completely analogous steps. It then remains to show that

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^p}(Y_{it+1}, Y_{it}, Y_{it-(2p-1)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-p}, X_i, A_i\right] = \pi_t^{y_1|y_1^p}(A_i, X_i)$$

To this end, it suffices to repeat the calculations employed in the induction argument but using this time

$$\mathbb{E}\left[\phi_{\theta_0}^{y_1|y_1^{p-1}}(Y_{it+1}, Y_{it}, Y_{it-(2p-2)}^{t-1}, X_i)|Y_i^0, Y_{i1}^{t-(p-1)}, X_i, A_i\right] = \pi_t^{y_1|y_1^{p-1}, Y_{it-(p-1)}}(A_i, X_i)$$

$$k_t^{y_1|y_1^p}(\theta) = \sum_{r=1}^{p} \gamma_r y_r + X_{it+1}'\beta$$

$$u_{t-(p-1)}(\theta) = \sum_{r=1}^{p} \gamma_r Y_{it-(r+p-1)} + X_{it-(p-1)}'\beta$$

$$w_t^{y_1|y_1^p}(\theta) = \left[1 - e^{(k_t^{y_1|y_1^p}(\theta) - u_{t-(p-1)}(\theta))}\right]^{y_p} \left[1 - e^{-(k_t^{y_1|y_1^p}(\theta) - u_{t-(p-1)}(\theta))}\right]^{1-y_p}$$