Asymptotic Properties of the Maximum Likelihood Estimator under Occasionally Observed States*

Alexandros Gilch† Gregor Reich‡ Ole Wilms§¶

October 2024

PRELIMINARY AND INCOMPLETE DRAFT

Abstract

Estimating Markov models with hidden state variables presents significant challenges because the likelihood involves a high-dimensional integral over the unobserved states. This complication renders the standard approach to prove the asymptotic properties of the likelihood-based estimator infeasible, because it relies on a log-transformation of the likelihood function. Moreover, the need to numerically approximate the integral in the likelihood function introduces an additional source of error in the estimation process. In this paper, we demonstrate how occasional observations of the hidden state restore the feasibility of the log-likelihood approach for establishing asymptotic properties, thereby extending existing results to general state spaces for the hidden state. Further, we show that, given consistency and asymptotic normality of the exact estimator, the desired properties can be extended to the estimator based on the approximated likelihood.

Keywords: maximum likelihood estimation, occasional state observations, recursive likelihood function integration, large sample properties

*We thank Björn Höppner and Jan Scherer for helpful comments and discussions. Gregor Reich gratefully acknowledge the financial support of Kenneth Judd, Senior Fellow at the Hoover Institution. Alexandros Gilch gratefully acknowledges financial support by the Collaborative Research Center Transregio 224, funded by the German Research Foundation (DFG).

†Institute of Finance and Statistics, University of Bonn, Adenauerallee 24-26, 53113 Bonn, Germany. Email: alexandros.gilch@uni-bonn.de

‡Tsumcor Research AG, Sonnenbergstrasse 74, 8603 Schwerzenbach, Switzerland. Email: gregor.reich@tsumcor.ch

§Department of Economics, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany. Email: ole.wilms@uni-hamburg.de

¶Department of Finance, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands.

1 Introduction

The estimation of structural dynamic economic models often encounters limited data availability, having certain state variables observed only intermittently across time periods. We refer to such state variables as *occasionally observed*.[1] In maximum likelihood estimation, when a realization of a state is not observed by the econometrician, but has (potentially) been relevant for the decision maker, it must be marginalized by integrating it out of the density of the data that induces the likelihood of the parameters to be estimated. However, if the occasionally observed states are serially correlated—as it is typically the case in realistic dynamic models—this integral tends to be high-dimensional, and cannot be decomposed analytically into smaller integrals. This results in two issues: First, the integral structure of the likelihood makes the application of a law of large numbers on the log-likelihood impossible; thus, even if the integral can be computed in closed form, the large sample properties of this *exact estimator* are not ex ante innate. Second, since the former assertion typically fails to hold in concrete applications, the integral has to be approximated numerically; consequently, the *approximate estimator* is subject to an additional source of error, making its asymptotic properties even more ambiguous.

For some special classes of models, the literature has developed the asymptotic theory for estimators based on integrated likelihoods that justify their application, most notably the hidden Markov models (HMMs) literature. However, while the latter has indeed established consistency and asymptotic normality for HMMs, it remains restrictive in three key aspects: First, it assumes complete unobservability of at least one state variable, and thus fails to fully exploit the available data in the presence of occasional observations. Second, it imposes limitations on the form of the transition density; specifically, HMMs rely on a conditional independence assumption, wherein the observed states at a given period t are independent of observations from other periods—particularly from $t-1$—given the unobserved state at period t. Third, the results typically rely on the restrictive assumption that the domain of the unobserved states is effectively compact, which is prohibitive for many economic models where random unobserved components are often drawn from an auto-regressive distribution with normal innovations whose support is the real line.

Gilch et al. (2024) address the three limitations by demonstrating how to incorporate occasional observations into the likelihood function, while accounting for potential endogeneity arising from the inter-dependency between observability and the realization of the state variables. To evaluate the resulting function numerically, they invoke a recursive likelihood integration method (RLI; originally due to Reich, 2018), thereby enabling likelihood-based inference in the presence of occasionally observed states. In this companion paper to Gilch et al. (2024), we provide the large sample properties of the proposed estimator, which justify its use theoretically. In particular, we exploit the presence of occasional observations to establish consistency and asymptotic normality of the exact estimator for general state domains, including the (full) real line, and for general transition densities under mild regularity requirements.[2] Moreover, since Gilch et al. (2024) apply numerical methods that are naturally subject to approximation error,

[1] Hall and Rust (2021) refer to such variables as *endogenously sampled*.

[2] Hall and Rust (2021) establish asymptotic results for a simulated method of moments estimator, applicable to a similarly general class of models compared to ours.

we apply and extend the proof strategies of Griebel et al. (2019) to demonstrate that also the approximate estimator remains both consistent and asymptotically normal.

Consistency and asymptotic normality of the maximum likelihood estimator (MLE) in hidden Markov models (HMMs) have been the subject of significant research. The foundational work of Baum and Petrie (1966) establishes these properties for HMMs with finite state spaces and introduces the influential "infinite-past" proof strategy, which serves as a cornerstone for much of the subsequent literature. Building on this, Leroux (1992) and Bickel et al. (1998) extend the consistency results to HMMs where the hidden state space remains finite, but the observation space is allowed to be general. Further advancements include the work of Douc et al. (2004), who study autoregressive HMMs, introducing an additional dependency channel from past observations. They demonstrate both consistency and asymptotic normality under the assumption that the hidden state space is compact and the observation space is general. Cappé et al. (2005) generalize these results by addressing HMMs with completely general state spaces for both the hidden and observed variables, though their results hinge on specific conditions on the transition probabilities, which may not always hold in practice. Douc et al. (2011) refine the understanding of consistency by removing the assumption of uniform positivity, i.e., de facto compactness of the state space, which was prevalent in earlier analyses.

We contribute to this literature by leveraging occasional observations of the hidden variable to eliminate the compactness assumption on the state space for establishing both the consistency and asymptotic normality of our estimator. Notably, we do not impose the specific structure of hidden Markov models, thereby extending existing results to address the large sample properties of general Markov models with occasionally observed states. Note that maximum likelihood estimation inherently assumes point identification of the analyzed model—and so do we. Restricting oneself to a certain class of models may allow the econometrician to apply non-likelihood based estimation approaches that circumvent this assumption. E.g., for dynamic discrete choice models, Kasahara and Shimotsu (2009) present conditions that imply non-parametric identifiability, while Berry and Compiani (2020) use an instrumental variables approach to show identification in dynamic models from the IO literature. For HMM, Cappé et al. (2005) discuss classes of transition densities for which identifiability can be derived and Douc et al. (2011) provide an information-theoretic argument, which yields identifiability under certain conditions.

Despite the extensive literature on the asymptotic properties of the exact likelihood estimator, relatively little attention has been given to how these properties carry over to estimators based on approximate likelihoods. Broadly, approximation methods for these integrals fall into two categories. The first consists of simulation-based approaches, such as particle filter methods, which simulate possible sequences of the unobserved states to construct a Monte Carlo approximation of the likelihood. For these methods, the established literature on simulated maximum likelihood provides starting points for a deeper theoretical analysis. The second category involves numerical integration methods, which are deterministic and highly efficient but inherently subject to the curse of dimensionality. Unlike simulation-based methods, these deterministic approximations exhibit a non-vanishing error when the number of integration nodes

is fixed. The approximation error decreases only as the accuracy of the approximation improves asymptotically. Importantly, this error does not only increase the variance of the estimator but introduces a bias that cannot be controlled when using a small number of approximation nodes. This situation calls for a dedicated analysis of estimators based on approximated likelihoods, as pioneered by Griebel et al. (2019).

We base our proofs on the presentation in Newey and McFadden (1994) for extremum estimators, and recent results by Griebel et al. (2019) for the estimators including the numerical approximation of an integral in their definition. A key feature of these findings are assumptions about regularity of the likelihood function and conditions regarding the search domain for the estimated parameters. We show how to use these conditions to obtain consistency and asymptotic normality for the approximate estimator of Markov models with occasionally observed state variables.

The remainder of this paper is organized as follows: In Section 2, we first present Markov models with occasionally observed state variables in a general framework and then introduce the recursive likelihood integration approach. In Section 3, we proof consistency and asymptotic normality of the exact estimator based on the frequency of the occasional observations. In Section 4, we provide conditions under which both properties also hold for the approximate estimator.

2 Markov models with occasionally observed states

In this section we present the general likelihood for Markov models with occasionally observed states and demonstrate the potential challenges for deriving large sample properties of the maximum likelihood estimator. For this, we first use a simplified setting with only two state variables—one that is always observed and one that is only observed occasionally. We show that the likelihood forms a high-dimensional integral over a non-standard domain. This simplified example allows us to demonstrate how the integral structure of the likelihood interferes with the standard loglikelihood approach for showing asymptotic properties of the maximum likelihood estimator. Afterwards, we generalize our notation to include an arbitrary number of state variables with general observation patterns. Note that in this section we closely follow the notation of the companion paper Gilch et al. (2024) to demonstrate the applicability of our results to the RLI estimator proposed in Gilch et al. (2024). Hence the presentation and derivation of the likelihood in this paper does not form a contribution in itself.

2.1 The likelihood in a simplified model with one-dimensional states

Consider a discrete-time Markov process $\{y_t, x_t\}$—possibly controlled, like in dynamic discrete choice models—with two one-dimensional state variables, $y_t, x_t \in \mathbb{R}$, and a parametric family of transition probability functions, $P(y_t, x_t | y_{t-1}, x_{t-1}; \theta)$. We want to estimate the model parameter θ using a maximum likelihood approach. In particular, we are interested in a case of limited data availability, where the variable y_t is observed for all periods $t \in \mathcal{T} \equiv \{1, \dots, T\}$ of the sample, whereas x_t is observed only at the times $t \in \bar{\mathcal{T}}$ with $\bar{\mathcal{T}} \subseteq \mathcal{T}$.

To introduce basic notation and the fundamental treatment of unobserved states, let us

first consider two counterfactual cases: Under full observability for both states x_t and y_t—i.e., $\bar{\mathcal{T}} = \mathcal{T}$—the (unconditional) likelihood function of the parameter vector θ reads

$$\begin{aligned} L^T(\theta) &= P(\{y_t, x_t\}_{t\in\mathcal{T}}; \theta) \\ &= P(y_1, x_1|\theta) \prod_{t=2}^{T} P(y_t, x_t|y_{t-1}, x_{t-1}; \theta), \end{aligned} \tag{1}$$

where $P(y_1, x_1|\theta)$ is the stationary distribution of x_t (if available). Conversely, if no state observations on x_t are available—i.e., $\bar{\mathcal{T}} = \emptyset$—the likelihood function forms an integral with respect to the unobserved state,

$$\begin{aligned} L^T(\theta) &= P(\{y_t\}_{t\in\mathcal{T}}; \theta) \\ &= \int \cdots \int_{\mathcal{S}_x^T} P(y_1, \tilde{x}_1|\theta) \prod_{t=2}^{T} P(y_t, \tilde{x}_t|y_{t-1}, \tilde{x}_{t-1}; \theta) \, \mathrm{d}(\tilde{x}_1, \ldots, \tilde{x}_T) \,. \end{aligned} \tag{2}$$

Here and in the following, we decorate any integration variable with a tilde; in (2), we write $\tilde{x}_t$ to clearly distinguish them from any data set element or state variable, x_t. Note that the overall dimensionality of the integral in (2) is proportional to the time horizon of the data, T. Thus, computing this integral constitutes a delicate task.

Suppose we have a single observation $x_{\bar{t}}$ at $\bar{t}$ that lies in the "interior" of $\mathcal{T}$—i.e., $1 < \bar{t} < T$ and $\bar{\mathcal{T}} = \{\bar{t}\}$. If we were to integrate the likelihood as in (2), the domain of integration in the likelihood function would read $\{(\tilde{x}_1, \ldots, \tilde{x}_T) \in \mathcal{S}_x^T : \tilde{x}_{\bar{t}} = x_{\bar{t}}\}$, which is no longer a full-dimensional subset of $\mathcal{S}_x^T$ (for general state spaces $\mathcal{S}_x$), and thus potentially creates ill-defined integrals. Therefore, we rewrite the integral to explicitly exclude the integration variable $\tilde{x}_{\bar{t}}$ and only integrate w.r.t. the unobserved states $\tilde{x}_t$ for $t \in \mathcal{T} \setminus \bar{\mathcal{T}}$:

$$\begin{aligned} L^T(\theta) = \int \cdots \int & P(y_1, \tilde{x}_1|\theta) \left(\prod_{t=2}^{\bar{t}-1} P(y_t, \tilde{x}_t|y_{t-1}, \tilde{x}_{t-1}; \theta) \right) P(y_{\bar{t}}, x_{\bar{t}}|y_{\bar{t}-1}, \tilde{x}_{\bar{t}-1}; \theta) \\ & \cdot P(y_{\bar{t}+1}, \tilde{x}_{\bar{t}+1}|y_{\bar{t}}, x_{\bar{t}}; \theta) \left(\prod_{t=\bar{t}+2}^{T} P(y_t, \tilde{x}_t|y_{t-1}, \tilde{x}_{t-1}; \theta) \right) \mathrm{d}(\tilde{x}_1, \ldots, \tilde{x}_{\bar{t}-1}, \tilde{x}_{\bar{t}+1}, \ldots, \tilde{x}_T) \,. \end{aligned} \tag{3}$$

In the companion paper, Gilch et al. (2024), we use the likelihood (3) to illustrate three key challenges that arise when estimating dynamic models with occasional state observations. First, observation of x_t may depend on the realization of the variable itself, meaning that whether or not we observe x_t could convey information about its value. Ignoring this dependency in constructing the likelihood can lead to biased estimators. Second, the integral in equation (3) is numerically challenging, as its dimension grows proportionally with the time horizon T, subjecting numerical quadrature methods to a curse of dimensionality. Third, due to the integral structure, the asymptotic properties of the likelihood estimator are not ex ante clear because the standard log-likelihood approach to proving these properties is infeasible. Additionally, the estimator based on the approximated likelihood integral also includes an approximation error, making its asymptotic properties even more difficult to characterize.

We discuss the first and second challenges in Gilch et al. (2024), so we will not explain them

in detail here. To summarize, the first challenge is addressed by incorporating information about the observation process—specifically, the mechanism determining whether x_t is observed or not—into the likelihood formulation. Regarding the second challenge, Gilch et al. (2024) introduces a generalized version of the recursive likelihood integration (RLI) method developed by Reich (2018). This approach effectively avoids the curse of dimensionality and provides a highly efficient deterministic approximation of the likelihood as is demonstrates in several applications from the finance and IO literature.

Although we do not discuss the challenges in detail, we introduce the mechanism and notation of the RLI method in this paper for two reasons. First, to our knowledge, it is the only practical implementation of a deterministic algorithm for computing the likelihood integral. Therefore, in order to demonstrate the relevance of our theoretical results we discuss how its fast convergence properties can actually improve estimation compared to simulation approaches. Second, for asymptotic normality, our proof requires approximations of the Jacobian and the Hessian with sufficient convergence rates. In Section 4.2, we show how the RLI method can provide these approximations.

The third challenge—the large-sample properties of the maximum likelihood estimator—can be addressed in two parts: first, demonstrating how occasional observations establish the asymptotic properties of the estimator based on the exact integral; and second, showing that the RLI approximation (or any similar approximator with comparable convergence properties) does not distort these properties in the limit. In the following sections, we illustrate both approaches using an example with a single observation.

To proceed, we first introduce the two estimators central to our analysis. The exact estimator, $\hat{\theta}_T$, is based on the exact likelihood L^T, which is generally not attainable because L^T does not admit a closed form,

$$\hat{\theta}_T = \operatorname*{argmax}_{\theta \in \Theta} L^T(\theta). \tag{4}$$

This estimator is indexed by T, the length of the time series, which we let approach infinity to establish its asymptotic properties. The approximate estimator, $\tilde{\theta}_{TN}$, is based on the approximated likelihood $\tilde{L}^{TN}$, defined formally in the next section,

$$\tilde{\theta}_{TN} = \operatorname*{argmax}_{\theta \in \Theta} \tilde{L}^{TN}(\theta). \tag{5}$$

It is indexed by both T (the time dimension) and N, representing the numbers of nodes used in each computation step of the approximation and, thus, representing the accuracy of the approximation. We formally define N in Section 4.1.

Failure of the standard loglikelihood approach. In the standard setup with fully observed x_t, asymptotic properties of the likelihood estimator $\hat{\theta}_T$ are obtained by taking the logarithm of the likelihood (1),

$$\log L^T(\theta) = \log P(y_1, x_1|\theta) + \sum_{t=2}^{T} \log P(y_t, x_t|y_{t-1}, x_{t-1}; \theta).$$

Taking the sample size T to infinity, consistency is derived using a law of large numbers and asymptotic normality follows from a central limit theorem.

In the case of occasional state observations, this approach is not applicable: Taking the logarithm of the likelihood (3) does not yield a sum over t summands, but rather a sum of only two summands, as the logarithm and the integral cannot be interchanged:

$$\begin{aligned}\log L^T(\theta) = \log \int \cdots \int P(y_1, \tilde{x}_1|\theta) \left(\prod_{t=2}^{\bar{t}-1} P(y_t, \tilde{x}_t|y_{t-1}, \tilde{x}_{t-1}; \theta) \right) P(y_{\bar{t}}, x_{\bar{t}}|y_{\bar{t}-1}, \tilde{x}_{\bar{t}-1}; \theta) \, \mathrm{d}(\tilde{x}_1, \ldots, \tilde{x}_{\bar{t}-1}) \\ + \log \int \cdots \int P(y_{\bar{t}+1}, \tilde{x}_{\bar{t}+1}|y_{\bar{t}}, x_{\bar{t}}; \theta) \left(\prod_{t=\bar{t}+2}^{T} P(y_t, \tilde{x}_t|y_{t-1}, \tilde{x}_{t-1}; \theta) \right) \mathrm{d}(\tilde{x}_{\bar{t}+1}, \ldots, \tilde{x}_T) \,.\end{aligned} \tag{6}$$

Of course, with a fixed number of observations (here: one), we do not obtain the infinite sum required to derive the desired properties of the estimator when taking T to infinity. Instead, it is only the dimension of the integral that becomes larger, and convergence is in our general framework—to the best of our knowledge—unclear. This makes asymptotic statements impossible, even if we were able to compute these integrals exactly.

However, in Section 3 we show, that it is possible to recover the asymptotical results known from many other log-likelihood-based estimators, if the number of occasional observations, $S \equiv |\bar{\mathcal{T}}|$, also tends to infinity as T grows. Then, these asymptotics can be derived based on the joint probability of all states between two observation periods, $P\left(\{y_t, x_t\}_{t=\bar{t}_i+1}^{\bar{t}_{i+1}} | y_{\bar{t}_i}, x_{\bar{t}_i}; \theta\right)$. Importantly, this approach rests on the assumption that the period of non-observation between two full observations of x_t is bounded.[3] In Section 3.1, we formalize this setup as Assumption A1 and discuss its implications in often encountered applications.

Error due to approximation As mentioned above, our likelihood cannot be evaluated analytically and is therefore approximated numerically by some function $\tilde{L}^{TN}$. Consequently, the estimator we are actually interested in is not the maximizer of (6), $\hat{\theta}_T$, but rather the maximizer of this approximated likelihood, $\tilde{\theta}_{TN}$. However, since $L^T \neq \tilde{L}^{TN}$ implies $\hat{\theta}_T \neq \tilde{\theta}_{TN}$, approximating the likelihood introduces an additional deterministic error—one that does not vanish as the sample size grows—into our estimator. This approximation error compounds the estimator's stochastic estimation error, affecting its asymptotic properties.

The consistency and asymptotic normality of $\tilde{\theta}_{TN}$ can be derived if both properties hold for $\hat{\theta}_T$ and if the error $||\hat{\theta}_T - \tilde{\theta}_{TN}||$ vanishes in the limit. With the first condition being adressed in Section 3, we cover the second condition in Section 4: The idea behind our proof of the second condition is to improve the accuracy of the approximation in proportion to the length of the time series, meaning that N, the number of approximation nodes, should increase alongside T, the sample length. By doing so, as $T \to \infty$, the approximate estimator converges to the exact estimator at the same rate at which the exact estimator converges to the true parameter. The pace at which N needs to increase to achieve this convergence rate depends on the convergence properties of the approximation method yielding $\tilde{L}^{TN}$; for example, a fast-converging method

[3]Here, full implies that in the case of a occasionally observed state vector x_t all components of this vector are observed at the same time.

like the RLI algorithm requires only a relatively gradual increase in N, proportional to T.

2.2 The likelihood and its approximation in the general model

We introduce a general notation to formulate the likelihood function of a Markov model with serially correlated states, where some (or all) of the model states are observed only occasionally. In fact, the notation presented below allows for arbitrary observations patterns both w.r.t. time and the state space dimension. Again, we follow closely the notation and presentation in the appendix of the companion paper Gilch et al. (2024), which establishes the RLI algorithm for occasionally observed states.

In contrast to the previous section, we consider observed variables Y_t and occasionally observed variables X_t jointly to allow for general observation patterns and simplify notation in this section. Hence, we consider a stochastic process $\{W_t\}_{t\in\mathbb{N}}$, where the random vector has support $\mathcal{S} \subseteq \mathbb{R}^d$, which we refer to as the "state space".[4] Note that we restrict our attention to continuous state variables here, as all concepts we present below have simple analogues in the discrete case.

We assume the Markov model explaining $\{W_t\}$ to define a parametric family of (conditional) distributions, which can be represented through probability density functions

$$P\left(W_t \mid \{W_s\}_{s<t}; \theta\right) = P\left(W_t \mid W_{t-1}; \theta\right)$$

with $\theta \in \Theta \subset \mathbb{R}^p$.[5] In order to precisely express the observation pattern of a dataset, we introduce some more notation: Let $w_{\tau_0} \equiv (w^i)_{i\in\tau_0}$ denote the sub-vector of states for some index set $\tau_0 \subseteq \tau \equiv \{1, \ldots, d\}$. Moreover, we write $\tilde{\tau}_0 \equiv \tau \setminus \tau_0$ for the complement of τ_0 w.r.t. τ, and we express the number of dimensions of w_{τ_0} using the cardinality operator $|\tau_0|$. Finally, note that if we write $(w_{\tau_0}, w_{\tilde{\tau}_0})$, we tacitly assume the elements to be re-ordered appropriately so that $(w_{\tau_0}, w_{\tilde{\tau}_0}) = w$, including the special cases $(w_\tau, w_\emptyset)$ and $(w_\emptyset, w_\tau)$.

This notation allows us to define the observation pattern of a dataset as follows: For an observation horizon $\{0, \ldots, T\}$, the set of index sets $\{\tau_t\}_{t=0}^T$, $\tau_t \subseteq \tau$, specifies which dimensions of the state vector w are observed at each point in time t,[6] and we denote the dataset by $\{w_{t,\tau_t}\}_{t=0}^T$. Note that in order to distinguish entries of the dataset from generic sub-vectors of states such as w_{τ_t}, we have equipped the former with another time subscript besides the index set. This notation also allows us to implicitly distinguish completely, never, and occasionally observed variables and thus ties it back into the context of the previous section: A completely observed variable w_{ti} has $i \in \tau_t$ for all $t \in \mathcal{T}$, an unobserved variable has $i \in \tilde{\tau}_t$ for all t and a variable is occasionally observed if neither holds. At each point in time t, the state realizations

[4] We abstract from the more general case which supposes time-heterogenous dimensionality of the state space in favor of a lighter notation. As the integration dimension will vary over time due to occasional observations of $W_t = w_t$ this extension is straight-forward.

[5] The density P_θ can, of course, be time-dependent, but we spare the additional index here, as our notation encompasses this feature—theoretically—through a deterministic, discrete state.

[6] Note that in the outline of this section, we use a single index set $\mathcal{T}$ to denote the points in time where an observation of a single state takes place. Here, each point in time has its own index set τ_t, specifying the dimensions of the state space which are observed at time t.

are an element of the subset

$$\mathcal{S}_t \equiv \{w \in \mathcal{S} : w_{\tau_t} = w_{t,\tau_t}\},$$

which "binds" the observed dimensions to the values from the dataset. Note, though, that not necessarily all elements in $\mathcal{S}_t$ have non-zero probability density. For the integration over the unobserved dimensions, we also need the projection of $\mathcal{S}_t$ to the lower-dimensional space where the unobserved dimensions live:

$$\tilde{\mathcal{S}}_t \equiv \left\{\tilde{w} \in \mathbb{R}^{|\tilde{\tau}_t|} : \tilde{w} = w_{\tilde{\tau}_t}, w \in \mathcal{S}_t\right\}.$$

We write $\tilde{\mathcal{S}}_t = \emptyset$ if $\tau_t = \tau$ and thus $\tilde{\tau}_t = \emptyset$. The (unconditional) likelihood of the model under observation regime $\{\tau_t\}_{t=0}^T$ reads

$$\begin{aligned} L_g^T(\theta) &\equiv L(\theta|\{w_{t,\tau_t}\}_{t=1}^T) \\ &= \int \cdots \int_{\times_{t=1}^T \tilde{\mathcal{S}}_t} \prod_{t=1}^T P\left(\tilde{w}_t, w_{t,\tau_t}|\tilde{w}_{t-1}, w_{t-1,\tau_{t-1}}; \theta\right) \mathrm{d}\tilde{w}_T \ldots \mathrm{d}\tilde{w}_1 \qquad (7) \\ &= \int \cdots \int_{\times_{t=1}^T \tilde{\mathcal{S}}_t} \prod_{t=1}^T g_t\left(\tilde{w}_t, \tilde{w}_{t-1}, \theta\right) \mathrm{d}\tilde{w}_T \cdots \mathrm{d}\tilde{w}_1 \qquad (8) \end{aligned}$$

and thus resembles the definition of L_g^T in Reich (2018). The functions $g_t : \tilde{\mathcal{S}}_t \times \tilde{\mathcal{S}}_{t-1} \times \Theta \to \mathbb{R}$ are defined by

$$g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta) \equiv \begin{cases} P\left(\tilde{w}_t, w_{t,\tau_t}|\tilde{w}_{t-1}, w_{t-1,\tau_{t-1}}; \theta\right) & \text{if } t > 1 \\ P\left(\tilde{w}_1, w_{1,\tau_1}|\theta\right) & \text{if } t = 1 \end{cases}$$

s.t. the dependence of the integrand on the data is implicitly given in the subscript t of g_t. Note that both $\tilde{\mathcal{S}}_t$ and $\tilde{w}_t$ can be empty if $\tau_t = \tau$, i.e., g_t, g_{t+1} are constant in $\tilde{w}_t$ and no integration w.r.t. $\tilde{w}_t$ takes place. Using the Markov structure of the model and standard regularity conditions for g_t,[7] a Fubini–Tonelli theorem (the concrete version of it depending on the nature of $\mathcal{S}$) justifies a recursive formulation of (8),

$$\begin{cases} \varphi_t^\theta \in \mathbb{R}_+ : \begin{cases} 1 & t > T \\ g_t\left(w_t|w_{t-1};\theta\right)\varphi_{t+1}^\theta & \tau_t = \tau \\ \int_{\tilde{\mathcal{S}}_t} g_t\left((\tilde{w}, w_{t,\tau_t})|w_{t-1};\theta\right) \cdot f_{t+1}^\theta(\tilde{w})\mathrm{d}^{|\overline{\tau}_t|}\tilde{w} & \text{otherwise} \end{cases} & \tau_{t-1} = \tau \\ f_t^\theta : \tilde{\mathcal{S}}_{t-1} \to \mathbb{R}_+, w \mapsto \begin{cases} 1 & t > T \\ g_t\left(w_t|(w, w_{t-1,\tau_{t-1}});\theta\right)\varphi_{t+1}^\theta & \tau_t = \tau \\ \int_{\tilde{\mathcal{S}}_t} g_t\left((\tilde{w}, w_{t,\tau_t})|(w, w_{t-1,\tau_{t-1}});\theta\right) \cdot f_{t+1}^\theta(\tilde{w})\mathrm{d}^{|\overline{\tau}_t|}\tilde{w} & \text{otherwise} \end{cases} & \text{otherwise,} \end{cases} \qquad (9)$$

[7] These follow from the fact that g_t is derived from a conditional p.d.f. which tend to be continuous and bounded in most economic applications.

and the final likelihood reads

$$L(\theta; \{w_{t,\tau_t}\}_{t=1}^{T}) = \begin{cases} g_t\left(w_1; \theta\right) \varphi_2^\theta & \tau_1 = \tau \\ \int_{\tilde{\mathcal{S}}_1} g_t\left((\tilde{w}, w_{1,\tau_1}); \theta\right) f_2^\theta(\tilde{w}) \mathrm{d}^{|\overline{\tau}_1|}\tilde{w} & \text{otherwise.} \end{cases} \tag{10}$$

While formulation (9) is exact, it is not practical for implementation purposes for the following reasons:

1. Actually evaluating the final likelihood—and thus evaluating either f_2^θ or φ_2^θ—would still require traversing a tree with $T-1$ levels and potentially infinitely many "knots" at each level; thus its computational complexity would explode.

2. No explicit use is made from the knowledge of the observations w_{t,τ_t} to determine the *conditional* distribution of $w_{\overline{\tau}_t}$.

To address issue 1, we introduce a mapping between two function spaces $\mathcal{B}_n$ and $\mathcal{P}_n$, whose elements are real functions of n-dimensional arguments, and with all elements in $\mathcal{P}_n$ having a complete representation through a countable set of parameters:

$$\mathcal{I}_n : \mathcal{B}_n \to \mathcal{P}_n, f \mapsto \hat{f},$$

where

$$f, \hat{f} : \mathbb{R}^n \supseteq \mathcal{D} \to \mathbb{R},$$

and with the norm $\|f - \hat{f}\|$ being "small" in the appropriate sense.

As indicated in issue 2, knowledge of w_{t,τ_t} can be used in many instances to obtain "high density regions" for $w_{\overline{\tau}_t}$ by conditioning its distribution on w_{t,τ_t}. This can often be exploited when numerically approximating the integrals in (9), e.g., by placing the nodes of quadrature rules accordingly. Therefore, we rewrite the relevant cases, conditioning the integrated probability densities on the observed states; note that in practice, this is not always possible.

Consequently, the final likelihood function recursion reads

$$\begin{cases} \hat{\varphi}_t^\theta \in \mathbb{R}_+ : \begin{cases} 1 & t > T \\ g_t\left(w_t|w_{t-1};\theta\right)\hat{\varphi}_{t+1}^\theta & \tau_t = \tau \\ \int_{\mathcal{S}} g_t\left(\tilde{w}|w_{t-1};\theta\right)\hat{f}_{t+1}^\theta(\tilde{w})\mathrm{d}^n\tilde{w} & \tau_t = \emptyset \\ g_t\left(w_{t,\tau_t}|w_{t-1};\theta\right) & \\ \quad \cdot \int_{\tilde{\mathcal{S}}_t} g_t\left(\tilde{w}|w_{t,\tau_t},w_{t-1};\theta\right)\hat{f}_{t+1}^\theta(\tilde{w})\mathrm{d}^{|\overline{\tau}_t|}\tilde{w} & \text{otherwise} \end{cases} & \tau_{t-1} = \tau \\ \hat{f}_t^\theta : \tilde{\mathcal{S}}_{t-1} \to \mathbb{R}_+, w \mapsto \begin{cases} 1 & t > T \\ g_t\left(w_t|(w,w_{t-1,\tau_{t-1}});\theta\right)\hat{\varphi}_{t+1}^\theta & \tau_t = \tau \\ \mathcal{I}_{|\overline{\tau}_{t-1}|}\Big(\int_{\mathcal{S}} g_t\left(\tilde{w}|(w,w_{t-1,\tau_{t-1}});\theta\right) & \\ \qquad \cdot \hat{f}_{t+1}^\theta(\tilde{w})\mathrm{d}^n\tilde{w}\Big) & \tau_t = \emptyset \\ \mathcal{I}_{|\overline{\tau}_{t-1}|}\Big(g_t\left(w_{t,\tau_t}|(w,w_{t-1,\tau_{t-1}});\theta\right) & \\ \qquad \cdot \int_{\tilde{\mathcal{S}}_t} g_t\left(\tilde{w}|w_{t,\tau_t},(w,w_{t-1,\tau_{t-1}});\theta\right) & \text{otherwise} \\ \qquad \cdot \hat{f}_{t+1}^\theta(\tilde{w})\mathrm{d}^{|\overline{\tau}_t|}\tilde{w}\Big) & \end{cases} & \text{otherwise,} \end{cases} \tag{11}$$

and the actual likelihood can be computed analogously to (10).

3 Large sample properties of the exact maximum likelihood estimator with occasional observations

In this section, we establish the consistency and asymptotic normality of the exact estimator $\hat{\theta}_T$ based on the true likelihood L^T. With occasional observations, we can decompose the likelihood integral over all unobserved states into a product of integrals, each covering only a single segment of unobserved states. Our primary assumption is that the mean (or maximum) duration between two periods in which all state variables are observed remains bounded—a condition that is trivial for a single occasionally observed state in most applications. Under this assumption, the proofs of consistency and asymptotic normality follow from basic stationarity and ergodicity arguments.

3.1 Assumptions on the observation pattern

The standard approach to proving the consistency of maximum likelihood estimators relies on three key assumptions:[8] the uniform convergence of the likelihood function to a limiting function, the continuity of this limiting function, and its unique maximization over a (compact) parameter space, which ensures the identification of the estimation problem. For asymptotic normality, similar conditions are required for the derivatives of both the likelihood and the limiting function. For a detailed discussion see Newey and McFadden (1994). In Appendix A.1, we restate their Theorems 2.1 and 3.1, which provide conditions for consistency and asymptotic normality of the maximum likelihood estimator, as Lemmas 1 and 2 in order to reference them

[8] This approach also applies to the broader class of extremum estimators.

in our proofs where necessary.

The main challenge of this approach is to identify the limit function to which the likelihood (3) converges as $T \to \infty$. The remaining two conditions—summarized in assumptions (i)-(iii) of the Lemmas 1 and 2—depend on the researcher's modeling choices, and we take them as given in this paper. However, the difficulty with establishing the limit function arises specifically from the structure of the likelihood and hence a general treatment of this condition is necessary. Standard econometric theory treating maximum likelihood estimation uses the loglikelihood to apply the law of large numbers that in turn provides uniform convergence of L_g^T. Given a dataset $w_1, \ldots, w_T$, a generic loglikelihood would take the form

$$\log L(\theta|w_1, \ldots, w_T) = \sum_{t=1}^{T} \log L(\theta|w_t) \tag{12}$$

s.t. taking T to infinity provides a limit function of the form $\ell(\theta) = \mathbb{E}[\log L(\theta|w_t)]$ independent of t. This approach is generally not applicable in our case with unobserved components $\tilde{w}_t = w_{t,\tilde{\tau}_t}$ of the stochastic process w_t: Integrating these out leads to the functional form (7) and hence an integral "around" the product of individual likelihood contributions of w_{t,τ_t}. Due to the nature of serial correlation of $\{X_t\}$, the product cannot be pulled out of the integral and hence taking the logarithm of L_g^T does not yield the sum form (12).

We utilize the *occasional complete observations* of the underlying state variable $W_{\bar{t}} = w_{\bar{t}}$ to make this approach feasible again. Occasional complete observations interrupt the progression of the unobserved series and "reset" it to start from an observed value. This allows us to separate the integral over the entire (asymptotically infinite) time series into smaller integrals over finite periods which are separated at complete observation times $\bar{t}$. We formalize the notion of occasional complete observations in the following assumption:

Assumption A1. *(Occasional complete observations)*
Assume that there is a number $\bar{T} > 0$ s.t. for all $t \geq 1$ there exists an integer $s \in \{0, ..., \bar{T}\}$ with $\tau_{t+s} = \tau$ (or analogously $\tilde{\tau}_{t+s} = \emptyset$).

Practically speaking, this means that the stochastic variable w_t is completely observed at least every $\bar{T}$ periods. It disallows intervals on non-observations of arbitrary length and disallows that components of w_t are never observed. We illustrate the consequences of the existence of such periods for the example $\tau_1 = \tau_{\bar{t}} = \tau_T = \tau$: This implies the state $W_{\bar{t}} = w_{\bar{t}}$ is completely observed and the integral over $\tilde{w}_{\bar{t}}$ is formally undefined as $\tilde{\mathcal{S}}_{\bar{t}} = \emptyset$. Abusing notation we let $\int_\emptyset h(v)dw = h(v)$ so that (7) is a valid expression even when including the case of a completely observed $w_{\bar{t}}$. Thus, the complete observation cancels integration w.r.t. $d\tilde{w}_{\bar{t}}$ and allows us to split up the integral at $w_{\bar{t}}$ into two integrals:

$$\begin{aligned}
&L\left(\theta \mid \{w_{t,\tau_t}\}_{t=1}^{T}\right) \\
&= \int_{\times_{t=2}^{\bar{t}-1} \tilde{\mathcal{S}}_t} \prod_{t=2}^{\bar{t}} g_t\left(\tilde{w}_t, \tilde{w}_{t-1}, \theta\right) d\tilde{w}_{\bar{t}-1} \cdots d\tilde{w}_1 \;\cdot\; \int_{\times_{\bar{t}+1}^{T-1} \tilde{\mathcal{S}}_t} \prod_{t=\bar{t}+1}^{T} g_t\left(\tilde{w}_t, \tilde{w}_{t-1}, \theta\right) d\tilde{w}_{T-1} \cdots d\tilde{w}_{\bar{t}+1} \\
&= L\left(\theta \mid w_1, w_{2,\tau_2}, ..., w_{\bar{t}}\right) L\left(\theta \mid w_{\bar{t}}, w_{\bar{t}+1,\tau_{\bar{t}+1}}, ..., w_T\right) .
\end{aligned} \tag{13}$$

We exploit this decomposition at complete observation points in the proof of Theorem 1. We further require two standard assumptions for inference with time series data.

Assumption A2. *The stochastic process $\{W_t\}_{t\in\mathbb{N}}$ is said to be stationary ergodic if the following conditions hold:*

- ***Stationarity**: For any $s, k, t_1, ..., t_k \in \mathbb{N}_0$ it holds that*

$$P_{\theta_0}(W_{t_1}, ..., W_{t_k}) = P_{\theta_0}(W_{t_1+s}, ..., W_{t_k+s})\,.$$

- ***Ergodicity**: For any two bounded functions $f : \mathbb{R}^{k+1} \to \mathbb{R}$, $g : \mathbb{R}^{\ell+1} \to \mathbb{R}$, and $k, l, i \in \mathbb{N}$ the process $\{W_t\}_{t\in\mathbb{N}}$ satisfies*

$$\lim_{n\to\infty} \Big| \mathbb{E}\left[f(W_i, \ldots, W_{i+k}) g(W_{i+n}, \ldots, W_{i+n+\ell})\right] - \mathbb{E}\left[f(W_i, \ldots, W_{i+k})\right] \mathbb{E}\left[g(W_{i+n}, \ldots, W_{i+n+\ell})\right] \Big| = 0\,.$$

Theorem 1. *Let $\{w_{t,\tau_t}\}_{t=1}^T$ be (partial) observations of a Markov process $\{W_t\}_{t=1}^T$ as defined in Section A.1 and suppose Assumptions A1 and A2 hold. Then, there exists a function $\ell(\theta; \hat{\tau}_{s^*})$ s.t. the loglikelihood $\ell^T(\theta) \equiv \log L_g^T(\theta)$ with $L_g^T(\theta)$ as defined in (7) satisfies a law of large numbers,*

$$\operatorname{plim}_{T\to\infty} \left| \frac{1}{T}\ell^T(\theta) - \ell(\theta; \hat{\tau}_{s^*}) \right| = 0\,.$$

Proof. See Appendix A.2. □

3.2 Consistency and asymptotic normality of the exact estimator

Theorem 1 shows that under Assumptions A1 and A2 we can find a limit function to which the loglikelihood ℓ^T converges in probability. This provides easy to verify conditions for the validity of assumption (iv) in Lemma 1. Together with the previous discussion of assumptions (i)-(iii), we summarize our findings in the following theorem.

Theorem 2. *Suppose Assumptions A1 and A2 hold and let $\ell(\theta; \hat{\tau}_{s^*})$ be defined as in Theorem 1. Furthermore assume that*

(i) $\ell(\theta; \hat{\tau}_{s^*})$ is uniquely maximized at $\theta_0 \in \Theta$,

(ii) the parameter space $\Theta \subseteq \mathbb{R}^p$ is compact, and

(iii) the functions g_t defined in (14) are continuous in $\theta \in \Theta$.

Then, the maximum likelihood estimator

$$\hat{\theta}_T = \operatorname*{argmax}_{\theta\in\Theta} L_g^T(\theta)$$

is consistent.

Proof. See Appendix A.2. □

We continue with proving asymptotic normality of $\hat{\theta}_T$ in a similar fashion by analyzing assumptions (i)-(vi) of Lemma 2: Assumption (i) repeats the conditions of Lemma 1 and is thus covered by the reasoning in the previous section. Similarly, (ii) is a direct result of choices regarding the estimation process and (iii) can be reduced to twice continuous differentiability of g_t w.r.t. θ. Assumptions (v) and (vi) follow from the existence of $H(\theta) \equiv \nabla_{\theta\theta}\ell(\theta; \hat{\tau}_{s^*})(\theta)$ and the assumption that θ_0 uniquely maximizes $\ell(\theta; \hat{\tau}_{s^*})$. Both of these statements follow from according modeling choices and from sufficient regularity of g_t. Convergence of $\ell^T(\theta)$ to $\nabla_{\theta\theta}\ell(\theta; \hat{\tau}_{s^*})(\theta)$ is achieved by a law of large numbers that follows from assumptions A1 and A2 in the same way as Theorem 1. Finally, assumption (iv) requires us to verify that the score function $\nabla_\theta L_g^T(\theta)$ satisfies a central limit theorem. Wooldridge (1994) states that the score $\{\nabla_\theta \ell(\theta \bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}})\}_{l\in\mathbb{N}}$ is a *martingale difference sequence* if the underlying time series $W_{tt\in\mathbb{N}}$ is *dynamically complete in distribution*. This is true in our case as $\{W_t\}_{t\in\mathbb{N}}$ denotes a Markov chain. Furthermore, for martingale difference sequences central limit theorems exist which satisfy assumption (iv) in Lemma 2. For a more detailed discussion of time series maximum likelihood and corresponding central limit theorems we refer to Wooldridge (1994). We summarize the above in the following theorem.

Theorem 3. *Suppose assumptions A1 and A2 hold and let $\ell(\theta; \hat{\tau}_{s^*})$ be defined as in Theorem 1. Furthermore, assume the assumptions of Theorem 2 are fulfilled and that*

(i) $\theta_0 \in \mathring{\Theta}$,

(ii) $g_t \in \mathcal{C}^2(\mathcal{N})$ *for a neighborhood* $\mathcal{N}$ *of* θ_0 *for all* $t \in \mathbb{N}$,

(iii) $H(\theta) \equiv \nabla_{\theta\theta}\ell(\theta; \hat{\tau}_{s^*})(\theta)$ *exists and is non-singular.*

Then, the maximum likelihood estimator $\hat{\theta}_T$ is asymptotically normal.

4 Large sample properties of the approximated maximum likelihood estimator with occasional observation

In the previous section, we have established that the exact maximum likelihood estimator is consistent and asymptotically normal. However, as the likelihood usually does not permit a closed-form solution it needs to be approximated, e.g., by the RLI method proposed in Gilch et al. (2024). The maximum likelihood estimator based on the approximate likelihood will differ from the exact likelihood estimator and hence, we must prove both large sample properties separately for the approximate estimator. To this end, we show that as the accuracy of the likelihood approximation improves, the approximate estimator converges to the exact one under weak regularity conditions. We provide rates for this improvement, ensuring that the approximate estimator inherits both consistency and asymptotic normality from the exact estimator.

4.1 Convergence properties of the RLI approximator

We begin by outlining the approximation properties of the generalized RLI approximation as introduced in Section 2.2.[9] Since, to the best of our knowledge, the RLI method is the only computationally feasible alternative to non-deterministic simulation methods, this is necessary to validate the applicability of our proof strategy in the subsequent section.

We use the following compact formulation of the recursive approximation scheme from Section 2.2,

$$\bar{g}_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta) \equiv g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta)\, \mathcal{I}_t \left(\hat{I}_t \left(\bar{g}_{t+1} \left(\cdot, \tilde{w}_t, \theta \right) \right) \right) \tag{14}$$

for $t = 1, ..., T-1$ and $\bar{g}_T \equiv g_T$, which includes all special cases regarding full observations ($\tau_t = \tau$ and/or $\tau_{t-1} = \tau$) implicitly. The approximated unconditional likelihood $\hat{L}_g^{TN}$ of θ given the observations $\{w_{t,\tau_t}\}_{t=1}^T$ is then given by

$$\hat{L}_g^{TN}(\theta) \equiv \hat{I}_1 \left(\bar{g}_1(\cdot, \theta) \right) . \tag{15}$$

The additional superscript N denotes the set $N = \{n_1, ..., n_T\}$ and indicates the approximation levels chosen for the interpolation steps $\mathcal{I}_t$ and the quadrature steps $\hat{I}_t$: Most economic models yield functional forms with sufficient regularity to assume the applicability of approximation rules with polynomial convergence rates, i.e., if the approximation uses n nodes, than the approximation error is of order $O(n^{-r})$ for some $r \geq 1$. Note that this notation even includes Monte Carlo simulation by setting $r = \frac{1}{2}$. Therefore, w.l.o.g. we choose interpolation methods $\mathcal{I}_t$, which use n_{It} interpolation nodes and have some polynomial convergence rate $O_{g_t}\left(n_{It}^{-r_{It}}\right)$, and quadrature rules $\hat{I}_t$, which use n_{Qt} quadrature nodes and have some polynomial convergence rate $O_{g_t}\left(n_{Qt}^{-r_{Qt}}\right)$.

For a given t the convergence properties of interpolation and quadrature may differ, i.e. $r_{It} \neq r_{Qt}$, generating approximation errors of different magnitude if the same number of nodes ($n_{Qt} = n_{It}$) were chosen for both methods. In these cases, the computational effort of approximation can be reduced by reducing the number of nodes used for the method, which converges quicker, i.e., for which the corresponding rate $r_{\cdot t}$ is larger. Similar to Reich (2018), we implement this idea by choosing balancing parameters ψ_t s.t. $n_{It} = n_t^{\psi_t}$ and $n_{Qt} = n_t^{1-\psi_t}$.

Given the respective rates r_{It}, r_{Qt}, the choices $\psi_t = \frac{r_{It}}{r_{It}+r_{Qt}}$ are optimal and yield the convergence rate $O_{g_t}(n_t^{-r_t})$ with $r_t = \frac{r_{It} r_{Qt}}{r_{It}+r_{Qt}}$. This definition is independent of the actual integration and interpolation dimension $|\tilde{\tau}_t|$ and the according choices of $\mathcal{I}_t$ and $\hat{I}_t$ as only the asymptotic convergence rates, given by $r_{\cdot t}$, enter in the total approximation error. The notation O_{g_t} indicates that the implied constant in the Landau notation is dependent on the integrand $\bar{g}_t$.

Approximation errors also accumulate over t. Therefore, it is optimal to choose n_t s.t. $n_t^{-r_t} = n_{t'}^{-r_{t'}}$ for all $t \neq t'$. This can be achieved by considering the smallest r_t (over $t = 1, ..., T$) and some series $\left\{ n_t^{(k)} \right\}_{k \in \mathbb{N}}$ with $n_t^{(k)} \to \infty$ for $k \to \infty$ and setting $n_{t'}^{(k)} = \left(n_t^{(k)} \right)^{r_t / r_{t'}}$. We denote the common convergence rate across $t = 1, ..., T$ by $O(n^{-r}) = O_{\bar{T},g}(n^{-r})$ with the constant depending on g_t and a maximal length $\bar{T}$ of the integrated time series, i.e. $T \leq \bar{T}$. Although we

[9] See Gilch et al. (2024) for a detailed treatment of the RLI estimator under occasional state observations.

will later take T to infinity, this condition is not problematic for us as it is fulfilled naturally by assumption A1. Finally, given at least r-times continuous differentiability and boundedness of all g_t and the appropriate choice of quadrature and interpolation methods, Proposition 2 of Reich (2018) shows the convergence rate

$$\left|\hat{L}_g^{TN}(\theta) - L_g^T(\theta)\right| = O(Tn^{-r}) \tag{16}$$

which is also applicable to our setup and the respective definition of $\hat{L}_g^{TN}$ examined in this paper.

4.2 Consistency and asymptotic normality of the approximate estimator

We conclude by providing two theorems that transfer the results on large sample properties of $\hat{\theta}_T$ to the approximate estimator $\tilde{\theta}_{TN}$. We base our proof on a similar approach described in Griebel et al. (2019) and cite their results as Lemma 3 and Lemma 4 in the appendix by providing easy to verify conditions under which these lemmas hold.

Theorem 4. *Let L_g^T and $\hat{L}_g^{TN}$ be defined as above. Assume that $\ell^T(\theta; \hat{\tau}_{s^*})$ is defined as in Theorem 1 and all assumptions of Theorem 2 are fulfilled. Assume further that*

$$\left|\hat{L}_g^{TN}(\theta) - L_g^T(\theta)\right| = O(Tn^{-r})$$

and that there exists $\delta > 0$ s.t. $P_\theta(\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}}) \geq \delta$ for all $\theta \in \Theta$, $\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}} \in \mathcal{S}_{\bar{t}_l^*} \times \cdots \times \mathcal{S}_{\bar{t}_{l+1}^*}$. Then the estimator $\tilde{\theta}_{TN}$ is consistent.

Proof. See Appendix A.2. □

Note that the assumption of lower boundedness for $P_\theta(\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}})$ is practically equivalent to compact support for the (occasionally) observed variables w_{t,τ_t}. This is a usual assumption in related simulation methods and also a frequent precondition for proving large sample properties in the HMM literature, where the state x_t is never observed. In particular, this makes the results in this section hold independently of the results in the previous section as we do not require occasional observations for the approximated likelihood but can rely on existing results for the exact estimator (see discussion in the introduction).

We complete this section with the proof of asymptotic normality of $\tilde{\theta}_{TN}$. For this purpose, we need to provide an intuition for the Jacobian $\nabla_\theta \hat{L}_g^{TN}$ and the Hessian $\nabla_{\theta\theta} \hat{L}_g^{TN}$ of the approximated likelihood and their convergence to $\nabla_\theta L_g^T$ and $\nabla_{\theta\theta} L_g^T$ respectively. The gradient of L_g^T is given by

$$\begin{aligned}
\nabla_\theta L_g^T(\theta) &= \int \nabla_\theta \left(\prod_{t=1}^{T} g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta) \right) \tilde{w}_T \cdots d\tilde{w}_1 \\
&= \sum_{t=1}^{T} \int \left(\prod_{s=1}^{T} g_s(\tilde{w}_s, \tilde{w}_{s-1}, \theta) \right) \frac{\nabla_\theta g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta)}{g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta)} \tilde{w}_T \cdots d\tilde{w}_1 \\
&= L_{\nabla_\theta g}^T(\theta)\,.
\end{aligned} \tag{17}$$

Having set $g_t \geq 0$, interchanging integration and differentiation is allowed if the g_t are integrable in $\tilde{w}_t, \tilde{w}_{t-1}$ for all θ and the gradients $\nabla_\theta g_t$ exist for all $\tilde{w}_t, \tilde{w}_{t-1}$ a.s. We define the likelihood gradient contributions

$$L^T_{g,\nabla g_t}(\theta) = \int \left(\prod_{s \neq t} g_s(\tilde{w}_s, \tilde{w}_{s-1}, \theta) \right) \nabla_\theta g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta) d\tilde{w}_T \cdots d\tilde{w}_1 \tag{18}$$

such that $\nabla_\theta L^T_g$ is given by

$$\nabla_\theta L^T_g(\theta) = \sum_{t=1}^{T} L^T_{g,\nabla g_t}(\theta) \,.$$

Note that $L^T_{g,\nabla g_t}$ is a function with image in $\mathbb{R}^p$ but all its components have the same structure as (8) just replacing g_s with $\nabla_\theta g_s$ for $s = t$ and otherwise leaving it the same. Now, the RLI approximation scheme can be applied separately to each $L^T_{g,\nabla g_t}(\theta)$ (considering each component of the p-dimensional function separately). Under similar assumptions on the partial derivatives $\frac{\partial}{\partial \theta_i} g_t$ for $i = 1, ..., p$ and $t = 1, ..., T$ as on g_t, Proposition 2 of Reich (2018) holds for the approximated gradient $\hat{L}^{TN}_{g,\nabla g_t}(\theta)$ with the same convergence rates. As mentioned in the paper, it is not necessary to bound g_t to below 1, but any finite bound suffices for the proof of Proposition 2. Hence, we also need to assume boundedness of $\nabla_\theta g_t$ in order to establish convergence to $L^T_{g,\nabla g_t}$. Summing up $\hat{L}^{TN}_{g,\nabla g_t}(\theta)$ over t provides the approximator

$$\hat{L}^{TN}_{\nabla_\theta g}(\theta) = \sum_{t=1}^{T} \hat{L}^{TN}_{g,\nabla g_t}(\theta)$$

for $\nabla_\theta L^T_g(\theta)$ which converges with rate (16) multiplied by T:

$$\left| \hat{L}^{TN}_{\nabla_\theta g}(\theta) - \nabla_\theta L^T_g(\theta) \right| = O\left(T^2 n^{-r}\right) \,.$$

Finally, we need to relate $\hat{L}^{TN}_{\nabla_\theta g}$ to the Jacobian $\nabla_\theta \hat{L}^{TN}_g$. The RLI scheme uses a combination of interpolation and quadrature methods. Quadrature methods $\hat{I}$ are linear in their argument (which is the integrand) by construction, hence $\nabla_\theta \hat{I}(f(\cdot, \theta)) = \hat{I}(\nabla_\theta f(\cdot, \theta))$ for some function $f(\cdot, \theta)$. The same holds for standard interpolation methods like Chebychev and Spline interpolation which usually solve a linear system of equation in evaluations of the interpolated function to obtain interpolation weights. Together this implies $\hat{L}^{TN}_{\nabla_\theta g} = \nabla_\theta \hat{L}^{TN}_g$. A similar approximation $\hat{L}^{TN}_{\nabla_{\theta\theta} g} = \nabla_{\theta\theta} \hat{L}^{TN}_g$ may be derived for $\nabla_{\theta\theta} L^T_g$ to obtain the convergence rate $O(T^3 n^{-r})$. We can now state asymptotic normality for $\tilde{\theta}_{TN}$:

Theorem 5. *Let L^T_g, $\hat{L}^{TN}_g$, and $\ell^T(\theta; \hat{\tau}_{s^*})$ be defined as above. Suppose that all assumptions of*

Theorem 3 are fulfilled. Furthermore, assume that

$$\begin{aligned}\left|\hat{L}_g^{TN}(\theta) - L_g^T(\theta)\right| &= O(Tn^{-r}),\\ \left|\hat{L}_{\nabla_{\theta g}}^{TN}(\theta) - \nabla_\theta L_g^T(\theta)\right| &= O\left(T^2 n^{-r}\right),\\ \left|\hat{L}_{\nabla_{\theta\theta g}}^{TN}(\theta) - \nabla_{\theta\theta} L_g^T(\theta)\right| &= O\left(T^3 n^{-r}\right),\end{aligned}$$

and that $\bar{g}_t$ is twice continuously differentiable in $\theta \in \Theta$ for all $t \in \mathbb{N}$. Then, $\tilde{\theta}_{TN}$ is asymptotically normal.

Proof. See Appendix A.2. □

The next corollary generalizes our results for non-compact state spaces.

Corollary 1. *To do.*

References

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563.

Berry, S. T. and Compiani, G. (2020). An instrumental variable approach to dynamic models. *SSRN Electronic Journal*.

Bickel, P. J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models. *The Annals of Statistics*, 26(4).

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer New York.

Douc, R., Moulines, E., Olsson, J., and van Handel, R. (2011). Consistency of the maximum likelihood estimator for general hidden markov models. *The Annals of Statistics*, 39(1).

Douc, R., Moulines, É., and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of Statistics*, 32(5).

Gilch, A., Lanz, A., Müller, P., Reich, G., and Wilms, O. (2024). "Small Data": Inference with Occasionally Observed States.

Griebel, M., Heiss, F., Oettershagen, J., and Weiser, C. (2019). Maximum approximated likelihood estimation. Submitted. Available as University of Bonn INS Preprint No. 1905.

Hall, G. and Rust, J. (2021). Estimation of endogenously sampled time series: The case of commodity price speculation in the steel market. *Journal of Econometrics*, 222(1):219–243.

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

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden markov models. *Stochastic Processes and their Applications*, 40(1):127–143.

Newey, W. K. and McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. In *Handbook of Econometrics*, pages 2111–2245. Elsevier.

Reich, G. (2018). Divide and conquer: Recursive likelihood function integration for hidden markov models with continuous latent variables. *Operations Research*, 66(6):1457–1470.

Wooldridge, J. M. (1994). Chapter 45 estimation and inference for dependent processes. In *Handbook of Econometrics*, pages 2639–2738. Elsevier.

A Appendix

A.1 Asymptotics for extremum and approximate estimators

We build our analysis on the standard presentation of results for large sample properties of extremum estimators by Newey and McFadden (1994) for the first stage. For the second stage, Griebel et al. (2019) provide a general framework for analysis of approximate estimators. In both papers $Q_k(\theta)$ denotes an objective function utilizing k data points which is maximized over some parameter space Θ by the *extremum estimator* $\hat{\theta}_k$. This corresponds to the exact likelihood L^T and the exact likelihood estimator $\hat{\theta}_T$. If $Q_k(\theta)$ cannot be computed analytically, it is approximated by $\hat{Q}_{kn}(\theta)$ using an approximation method with n nodes and the approximate estimator $\hat{\theta}_{kn}$ is its maximizer. This corresponds to the approximated likelihood $\tilde{L}^{TN}$ and the approximate likelihood estimator $\tilde{\theta}_{TN}$.

For reference, we cite Theorems 2.1 and 3.1 from Newey and McFadden (1994) for asymptotics of the exact estimator:

Lemma 1. *Assume that there is a function $Q(\theta)$ such that*

(i) $Q(\theta)$ is uniquely maximized at θ_0,

(ii) the parameter space Θ is compact,

(iii) $Q(\theta)$ is continuous,

(iv) the function $Q_k(\theta)$ converges uniformly in probability to $Q(\theta)$.

Then, $\hat{\theta}_k$ is a consistent estimator of θ_0, i.e. $\operatorname{plim}_{k\to\infty} \hat{\theta}_k = \theta_0$.

Lemma 2. *Suppose that there exists a function $Q(\theta)$ s.t.*

(i) the assumptions of Lemma 1 hold,

(ii) $\theta_0 \in \mathring{\Theta}$,

(iii) $Q_k(\theta) \in \mathcal{C}^2(\mathcal{N})$ for a neighborhood $\mathcal{N}$ of θ_0,

(iv) $\sqrt{k}\nabla_\theta Q_k(\theta_0) \xrightarrow{d} N(0, \Sigma)$,

(v) there exists a function $H(\theta)$ which is continuous at θ_0 s.t.

$$\sup_{\theta\in\mathcal{N}} ||\nabla_{\theta\theta} Q_k(\theta) - H(\theta)|| \xrightarrow{p} 0,$$

(vi) $H \equiv H(\theta_0)$ is non-singular.

Then $\hat{\theta}_k$ is asymptotically normal, i.e. $\sqrt{k}(\hat{\theta}_k - \theta_0) \xrightarrow{d} N(0, H^{-1}\Sigma H^{-1})$.

For the approximate estimator, we continue by citing Theorem 2 and 4 from Griebel et al. (2019):

Lemma 3. *Assume that there exists a function $Q(\theta)$ such that*

(i) the assumptions of Lemma 1 are fulfilled, and

(ii) $\hat{Q}_{kn}(\theta)$ *converges uniformly in probability to* $Q_k(\theta)$*, i.e.*

$$\text{plim}_{k\to\infty} \sup_{\theta\in\Theta} \left|\hat{Q}_{kn}(\theta) - Q_k(\theta)\right| = 0\,.$$

Then, $\hat{\theta}_{kn}$ *is a consistent estimator of* θ_0*, i.e.* $\text{plim}_{k\to\infty}\hat{\theta}_{kn} = \theta_0$*.*

Lemma 4. *Assume that there exists a function* $Q(\theta)$ *s.t.*

(i) the assumptions of Lemma 2 are fulfilled,

(ii) condition (ii) of Lemma 3 holds,

(iii) $\hat{Q}_{kn}(\theta) \in \mathcal{C}^2(\mathcal{N})$ *almost surely,*

(iv) $\text{plim}_{k\to\infty}\sqrt{k}\sup_{\theta\in\Theta}||\nabla_\theta\hat{Q}_{kn}(\theta) - \nabla_\theta Q_k(\theta)|| = 0$*,*

(v) $\text{plim}_{k\to\infty}\sup_{\theta\in\Theta}||\nabla_{\theta\theta}\hat{Q}_{kn}(\theta) - \nabla_{\theta\theta}Q_k(\theta)|| = 0$*.*

Then $\hat{\theta}_{kn}$ *is asymptotically normal with the same limit distribution as* $\hat{\theta}_k$*.*

A.2 Proofs for Sections 3 and 4

This section contains the proofs for Theorems 1–5.

A.2.1 Proof of Theorem 1

Proof. **Step 1**: We redefine the set $\bar{\mathcal{T}} \equiv \{t \in \mathcal{T} \mid \tau_t = \tau\}$ to be the subset of periods in which w_t is completely observed and let $S \geq 1$ s.t. $S+1 = |\bar{\mathcal{T}}|$ (note that $1 \in \bar{\mathcal{T}}$ is always assumed).[10] By Assumption A1, at least every $\bar{T}$-th period features a completely observed w_t hence $S \geq \frac{T}{\bar{T}}$. We number the periods of complete observation in an ascending order, $1 = \bar{t}_0 < \bar{t}_1 < ... < \bar{t}_S$ and assume w.l.o.g. that the last period T also features a completely observed w_T, i.e. $\bar{t}_S = T$. Finally, the interval lengths are given by $s_k \equiv \bar{t}_k - \bar{t}_{k-1}$ for $k = 1, ..., S$ such that by definition $\sum_{k=1}^{S} s_k = T$ and by A1 $s_k \leq \bar{T}$ for all k.

Step 2: Take the logarithm of L_g^T to obtain the loglikelihood $\ell^T = \log L_g^T$. Using (8) and

[10] This is consistent with our previous definition of $\bar{\mathcal{T}}$ which had $d_x = 1$ and thus any observation was equivalent to a complete observation.

the same argument as for (13) we get the decomposition into likelihood contributions,

$$\begin{aligned}
\ell^T(\theta) &= \ell\left(\theta \mid \{w_{t,\tau_t}\}_{t=1}^T\right) \\
&= \log L\left(\theta \mid \{w_{t,\tau_t}\}_{t=1}^T\right) \\
&= \log \int_{\times_{t=1}^T \tilde{\mathcal{S}}_t} \prod_{t=1}^T g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta) d\tilde{w}_T \cdots d\tilde{w}_1 \\
&= \log \prod_{k=0}^{S-1} \int_{\times_{t=\bar{t}_k+1}^{\bar{t}_{k+1}-1} \tilde{\mathcal{S}}_t} \prod_{t=\bar{t}_k+1}^{\bar{t}_{k+1}} g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta) d\tilde{w}_{\bar{t}_{k+1}-1} \cdots d\tilde{w}_{\bar{t}_k+1} \\
&= \sum_{k=0}^{S-1} \ell\left(\theta \mid \{w_{t,\tau_t}\}_{t=\bar{t}_k}^{\bar{t}_{k+1}}\right). \qquad (19)
\end{aligned}$$

Each likelihood contribution utilizes the set of observations $\{w_{t,\tau_t}\}_{t=\bar{t}_k}^{\bar{t}_{k+1}}$ s.t. all such sets only overlap at the complete observations $w_{\bar{t}_k,\tau_{\bar{t}_k}}, w_{\bar{t}_{k+1},\tau_{\bar{t}_{k+1}}}$.

Step 3: From $S \geq \frac{T}{\bar{T}}$ it follows that $S \to \infty$ if $T \to \infty$. Together with $s_k \leq \bar{T}$ this implies that there exists s^* s.t. $s_k = s^*$ for infinitely many k. We assumed that $\mathcal{S}$ is of fixed dimension $d < \infty$, hence for every interval of length s^* there are up to $(s^*-1)d$ single variables w_{ti}, $t = \bar{t}_k + 1, ..., \bar{t}_k + s^* - 1$ that may or may not be observed by the econometrician. As for the existence of s^* it is easy to see that there is at least one set $\hat{\tau}_{s^*} = \{(s,i) \mid 1 \leq s \leq s^*-1, 1 \leq i \leq d,\}$ s.t. given $S \to \infty$ for infinitely many $\bar{t}_k$ the variable $w_{\bar{t}_k+s,i}$ is observed if and only if $(s,i) \in \hat{\tau}_{s^*}$. In other words, we first claim that among all possible lengths of intervals $s_k \leq S$ there is at least one, denoted by s^*, that must occur infinitely many times if T tends to infinity. Secondly, there are only finitely many patterns of occasional observations of w_{ti} in each of these intervals of length s^*, so for $T \to \infty$ (i.e. $S \to \infty$) at least one of these patterns (characterized by $\hat{\tau}_{s^*}$) must realize infinitely many times. We can now define the subset

$$\bar{\mathcal{T}}^* \equiv \left\{\bar{t}^* \in \bar{\mathcal{T}} \mid \bar{t}^* + s \in \bar{\mathcal{T}}, x_{\bar{t}^*+s,i} \text{ observed iff. } (s,i) \in \hat{\tau}_{s^*}\right\} \subseteq \bar{\mathcal{T}} \subseteq \mathcal{T}$$

and number its elements in ascending order $\bar{t}_1^* < \bar{t}_2^* < ... < \bar{t}_{S^*}^*$. In particular, $S^* = |\bar{\mathcal{T}}^*|$. This yields the *homogenous loglikelihood function*

$$\ell^T(\theta; \hat{\tau}_{s^*}) = \sum_{l=1}^{S^*} \ell\left(\theta \middle| \{w_{t,\tau_t}\}_{t=\bar{t}_l^*}^{\bar{t}_{l+1}^*}\right). \qquad (20)$$

Step 4: $\ell^T(\theta; \hat{\tau}_{s^*})$ is homogenous in the sense that all likelihood contributions contain information on exactly the same variables and therefore have the exact same structure regarding the conditional probabilities P_θ. We define a compact notation

$$\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}} \equiv \{w_{t,\tau_t}\}_{t=\bar{t}_l^*}^{\bar{t}_{l+1}^*}$$

for the sets of observations between complete observations $\bar{t}_l^*$ and $\bar{t}_{l+1}^*$ and rewrite (20) as

$$\ell^T(\theta; \hat{\tau}_{s^*}) = \sum_{l=1}^{S^*} \log P_\theta(\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}}) \,.$$

By Assumption A2, the time series $\{\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}}\}_{l \in \mathbb{N}}$ is also stationary and ergodic and hence by the Ergodic theorem the likelihood $\ell^T(\theta; \hat{\tau}_{s^*})$ fulfills a weak law of large numbers

$$\text{plim}_{S^* \to \infty} \left| \frac{1}{S^*} \ell^T(\theta; \hat{\tau}_{s^*}) - \mathbb{E}\left[\log P_\theta(\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}})\right] \right| = 0 \,.$$

Step 5: If s^* and $\hat{\tau}_{s^*}$ are unique, then all other interval lengths $s \leq \bar{T}$ and observation patterns $\hat{\tau}_s$ for intervals of length s only appear finitely many times. This implies $\ell^T(\theta) \to \ell^T(\theta; \hat{\tau}_{s^*})$ and $\frac{S^*}{T} \to 1$ as $T \to \infty$ and hence $\frac{1}{T}\ell^T$ also converges to

$$\ell(\theta; \hat{\tau}_{s^*}) \equiv \mathbb{E}\left[\log P_\theta(\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}})\right] \tag{21}$$

in probability where $\ell(\theta; \hat{\tau}_{s^*})$ is independent of l by Assumption A2. This result can be extended to the case with multiple $\hat{\tau}_{s^*}$ (with possibly different s^*) by determining the relative frequencies $\alpha_{\hat{\tau}_{s^*}}$ of the individual $\hat{\tau}_{s^*}$ w.r.t. each other. Then, the limit function is the weighted mean

$$\ell(\theta; \hat{\tau}_{s^*}) \equiv \sum_{\hat{\tau}_{s^*}} \alpha_{\hat{\tau}_{s^*}} \, \mathbb{E}\left[\log P_\theta(\bar{w}_{\bar{t}_l^*, \hat{\tau}_{s^*}})\right]$$

with $\sum_{\hat{\tau}_{s^*}} \alpha_{\hat{\tau}_{s^*}} = 1$. Both sums are finite as there are at most $\sum_{s^* \leq \bar{T}} 2^{d(s^*-1)}$ possible different observation patterns under Assumption A1. □

A.2.2 Proof of Theorem 2

Proof. We have defined ℓ^T as logarithm of L_g^T, hence $L_g^T = \exp \ell^T$. Theorem 1 gives us a law of large numbers for $\frac{1}{T}\ell^T$ which yields the following law of large numbers for L_g^T

$$\text{plim}_{T \to \infty} \left| L_g^T(\theta)^{\frac{1}{T}} - \exp\left(\ell(\theta; \hat{\tau}_{s^*})\right) \right| = 0$$

due to continuity of exp. Monotonicity of exp implies that L_g^T is also uniquely maximized by θ_0. Put together, and considering the discussion of Assumptions (i)-(iii) of Lemma 1 at the beginning of Section 3, consistency of $\hat{\theta}_T$ follows directly from Lemma 1 and Theorem 1. □

A.2.3 Proof of Theorem 4

Proof. We show consistency by proving that the Assumptions (i) and (ii) in Lemma 3 hold for Q_k and $\hat{Q}_{kn}$. Assumption (i) was already discussed in Section 3, hence it remains to show (ii). Instead of using L_g^T we use ℓ^T for this purpose and define its approximator

$$\hat{\ell}^{TN}(\theta) = \log \hat{L}_g^{TN}(\theta) = \sum_{k=0}^{S-1} \log \hat{L}_g^{s_k N}(\theta)$$

with S and s_k as in the proof of Theorem 1 and $\hat{L}_g^{s_k N}(\theta)$ being the RLI approximator of $\ell\left(\theta \mid \{w_{t,\tau_t}\}_{t=\bar{t}_k}^{\bar{t}_{k+1}}\right)$ from equation (19). Focusing on the case with unique s^* and $\hat{\tau}_{s^*}$, we get that $s_k = s^*$ and $\{w_{t,\tau_t}\}_{t=\bar{t}_k}^{\bar{t}_{k+1}} = \bar{w}_{\bar{t}_l^*,\hat{\tau}_{s^*}}$ almost always and hence asymptotically

$$\hat{\ell}^{TN}(\theta) \to^T \hat{\ell}^{TN}(\theta; \hat{\tau}_{s^*}) = \sum_{l=1}^{S^*} \log \hat{L}_g^{s^* N}(\theta | \bar{w}_{\bar{t}_l^*,\hat{\tau}_{s^*}}) \,.$$

Here $\hat{L}_g^{s^* N}(\theta|\bar{w}_{\bar{t}_l^*,\hat{\tau}_{s^*}})$ is the approximation of $P_\theta(\bar{w}_{\bar{t}_l^*,\hat{\tau}_{s^*}})$. We can now follow the proof of Theorem 7 in Griebel et al. (2019) to show

$$\lim_{T\to\infty} P\left(\sup_{\theta\in\Theta}\left|\frac{1}{T}\hat{\ell}^{TN}(\theta) - \frac{1}{T}\ell^T(\theta)\right| > \varepsilon\right) = 0\,.$$

We can use the lower bound on $P_\theta(\bar{w}_{\bar{t}_l^*,\hat{\tau}_{s^*}})$ to apply a mean value theorem and obtain

$$\begin{aligned}\sup_{\theta\in\Theta}\left|\frac{1}{T}\hat{\ell}^{TN}(\theta) - \frac{1}{T}\ell^T(\theta)\right| &\leq \sup_{\theta\in\Theta}\sup_{\bar{w}_{\bar{t}_l^*,\hat{\tau}_{s^*}}} \frac{1}{\delta}\left|\hat{L}_g^{s^* N}(\theta|\bar{w}_{\bar{t}_l^*,\hat{\tau}_{s^*}}) - P_\theta(\bar{w}_{\bar{t}_l^*,\hat{\tau}_{s^*}})\right| \\ &\leq O(s^* n^{-r}) \end{aligned} \tag{22}$$

with the convergence rate from (16). The supremum over θ is gained from compactness of Θ and continuity of g_t in θ. Note that any increasing function $n = n(T)$ is sufficient to let the term (22) decrease to 0 as $T \to \infty$. This proves Assumption (iii) of Lemma 3 and hence consistency of $\tilde{\theta}_{TN}$ (as maximizer of $\hat{\ell}^{TN}$ and thus also of $\hat{L}_g^{TN}$).

□

A.2.4 Proof of Theorem 5

Proof. We show asymptotic normality by proving Assumptions (i)-(v) of Lemma 4: Both (i) and (ii) are given by assumption and are discussed in the proofs of Theorems 3 and 4. Twice differentiability of $\hat{L}_g^{TN}$ follows from twice differentiability of $\bar{g}_t$. This is an innocuous assumption as most economic models (and thus the resulting likelihood functions) feature smooth functions. Additionally, standard interpolation functions also require and return at least twice continuously differentiable inputs and outputs respectively in order to achieve polynomial convergence rates which we require anyway.

Assumptions (iv) and (v) follow from the convergence rates of $\hat{L}_{\nabla_\theta g}^{TN}$ and $\hat{L}_{\nabla_{\theta\theta} g}^{TN}$: We plug these into the respective conditions (iv) and (v) and obtain the requirements $O(T^{2.5} n^{-r}) \to 0$ and $O(T^3 n^{-r}) \to 0$ for $T \to \infty$. Practically, the longer sample is associated with more summands in the chain rule for the Jacobian and Hessian, as can be seen in Equation (17). Each summand adds an individual independent error term of order $O(Tn^{-r})$ or $O(T^2 n^{-r})$, respectively. To counteract this accumulation of errors, the approximation itself has to improve as $T \to \infty$. However, if the approximation has high polynomial convergence order r, then this can be achieved by only a moderately fast growing function $n = n(T)$. In fact, any function $n(T)$ of at least rate $T^{(3+\varepsilon)/r}$ with $\varepsilon > 0$ is sufficient for this purpose. □