Supplement to "Hidden Rust Models"

Benjamin Connault*

May 2016

1 Appendix for section 3: Identification

1.1 Proof of Theorem 1: generic identification structure

First, we prove the existence part of Theorem 1, i.e., that there is some exceptional region Λ_e satisfying conditions (i) and (ii). The uniqueness part (the fact that there is a unique such exceptional region minimal for the inclusion) will follow easily.

We will obtain the $\bar{\lambda}$-generic structure of the λ-solution set of $F\left(\lambda, \bar{\lambda}\right) = 0$ from the structure of the joint $\left(\lambda, \bar{\lambda}\right)$-solution set of $F\left(\lambda, \bar{\lambda}\right) = 0$.

Let $\mathbb{A} = \mathbb{A}^{2d_\lambda}(\mathbb{C})$ be the joint affine space for $\left(\lambda, \bar{\lambda}\right)$ and let W be the zero set of $F\left(\lambda, \bar{\lambda}\right)$ jointly in both variables:

$$W = V(F) = \left\{\left(\lambda, \bar{\lambda}\right) \in \mathbb{A} \middle| F\left(\lambda, \bar{\lambda}\right) = 0\right\}$$

Let $\bar{\mathbb{A}} = \mathbb{A}^{d_\lambda}(\mathbb{C})$ be the affine space for $\bar{\lambda}$ only and $\pi : W \to \bar{\mathbb{A}}$ the restriction of the coordinate projection $\pi\left(\lambda, \bar{\lambda}\right) = \bar{\lambda}$ to W. For any $\bar{\lambda}^\star$, write $W\left(\bar{\lambda}^\star\right)$ for the fiber of π at $\bar{\lambda}^\star$:

$$W\left(\bar{\lambda}^\star\right) = \pi^{-1}\left(\bar{\lambda}^\star\right) = \left\{\left(\lambda, \bar{\lambda}^\star\right) \in \mathbb{C}^3 \times \{\bar{\lambda}^\star\} \middle| F\left(\lambda, \bar{\lambda}^\star\right) = 0\right\}$$

*University of Pennsylvania, Department of Economics, connault@econ.upenn.edu

Thus, $W\left(\bar{\lambda}^\star\right)$ is the set of solutions in λ to the system of equations $F\left(\lambda, \bar{\lambda}^\star\right) = 0$ at a specific $\bar{\lambda}^\star \in \bar{\mathbb{A}}$: this is exactly the system that needs to be solved for identification analysis.

An equivalent statement of the existence part of Theorem 1 is that $W\left(\bar{\lambda}^\star\right)$ has a constant structure for $\bar{\lambda}^\star \in \bar{U}$ where $\bar{U}$ is a (Zariski) open dense subset of $\bar{\mathbb{A}}$. To see that this is an equivalent statement, let $\bar{F}$ be a non-zero system of polynomials such that $\Lambda_e = V\left(\bar{F}\right)$ as in Theorem 1. Remember that the zero-sets form the closed sets of the Zariski topology on $\bar{\mathbb{A}}$: saying that a property holds outside of $V\left(\bar{F}\right)$ is equivalent to saying that it holds on an open set $\bar{U}$, the complement of $V\left(\bar{F}\right)$. $\bar{U}$ is non-empty because $\bar{F}$ is not $\{0\}$. Any non-empty open set is dense in the Zariski topology. Thus, the statements are indeed equivalent.

To prove this equivalent statement we relate the generic structure of $W\left(\bar{\lambda}^\star\right)$ to the generic structure of W, the "joint" zero-set. Consider the decompositions in irreducible components of W and of $W\left(\bar{\lambda}^\star\right)$, for any $\bar{\lambda}^\star$:

$$W = \bigcup_{m=1}^{M} V_m \qquad \text{and} \qquad W\left(\bar{\lambda}^\star\right) = \bigcup_{j=1}^{J\left(\bar{\lambda}^\star\right)} V_j\left(\bar{\lambda}^\star\right)$$

We can use Theorem A.14.10 p. 349 from Sommese and Wampler (2005), adapted to our context:

Theorem: Theorem A.14.10 in Sommese and Wampler (2005)
There is a (Zariski) open dense set $\bar{U} \subset \bar{\mathbb{A}}$ such that for any $\bar{\lambda}^\star \in \bar{U}$, and any $1 \leq m \leq M$, if V_m is an irreducible component of W of dimension d_m, then $V_m \cap W\left(\bar{\lambda}^\star\right)$ is the union of a fixed number n_m (n_m independent of $\bar{\lambda}^\star$) of irreducible components $V_j\left(\bar{\lambda}^\star\right)$ of $W\left(\bar{\lambda}^\star\right)$ of dimension $d_m - d_\lambda$.

In particular, for any $\bar{\lambda}^\star \in \bar{U}$, $W\left(\bar{\lambda}^\star\right)$ has a fixed number of irreducible components of each dimension, which proves the existence statement of Theorem 1.

Note that $\bar{F}$ such that $\Lambda_e = V\left(\bar{F}\right)$ could include polynomials with coefficients in $\mathbb{C}$. In fact, we can do better and show that $\Lambda_e \cap \mathbb{R}^{d_\lambda}$ is the zero-set of a finite number of real polynomials. Indeed, $\Lambda_e \cap \mathbb{R}^{d_\lambda}$ is a closed subset of $\mathbb{R}^{d_\lambda}$ with the subtopology

inherited from $\mathbb{A}^{d_\lambda}(\mathbb{C})$, which coincides with $\mathbb{A}^{d_\lambda}(\mathbb{R})$.

Turning to the uniqueness part of Theorem 1, consider the union Λ_g of all $\bar{U}$, such that the existence statement holds for $\bar{U}$. Λ_g is open and dense. Λ_g satisfies the existence statement and contains by definition all the open dense sets that do, i.e., Λ_g is maximal for the inclusion.

1.2 Proof of other results

Proof of Corollary 1. This is a particular case of Theorem 1 where the generic identification structure has n_{ls} irreducible components of dimension zero and no component in other dimensions. □

Proof of Corollary 2. (i) Λ_g is Zariksi open in $\mathbb{C}^{d_\lambda}$ implies Λ_g is Euclidean open in $\mathbb{C}^{d_\lambda}$ implies $\Lambda_g \cap [0,1]^{d_\lambda}$ is Euclidean open in $[0,1]^{d_\lambda}$ (subspace topology). For the Euclidean topology, Λ_g is dense in $\mathbb{R}^{d_\lambda}$ implies Λ_g is dense in $]0,1[^{d_\lambda}$ implies Λ_g is dense in $[0,1]^{d_\lambda}$. Thus Λ_g is Euclidean open dense in $[0,1]^{d_\lambda}$.
(ii) Λ_e has Lebesgue measure zero in $\mathbb{R}^{d_\lambda}$ and thus has Lebesgue measure zero in $[0,1]^{d_\lambda}$. □

Proof of Corollary 3. This is a direct consequence of Theorem 2. □

1.3 Identification analysis in a toy model

This subsection carries a detailed identification analysis in a toy model. The model retains all the interesting features of a hidden Rust model, except for the dynamic aspects. It is a good opportunity to illustrate:

- General algebraic-geometric concepts such as the decomposition in irreducible components of a Zariski topological set.
- The mechanism behind Theorem 1.
- The advantage of automatic methods to compute the generic identification structure as well as the minimal exceptional region.

Consider an unobserved discrete random variable $X \in \{a, b\}$ and an observed discrete random variable $Y \in \{0, 1, 2, 3\}$. $X = b$ with probability p and $X = a$ with

probability $1-p$. If $X=a$, Y is binomial $B(3,p_a)$ — the sum of three biased coin flips — and if $X=b$, Y is binomial $B(3,p_b)$.

In an identification context, we ask if the statistical parameter $\lambda=(p,p_a,p_b)$ is identified from the distribution of Y given by:

$$\begin{aligned}
\mathbb{P}(Y=0) &= (1-p)(1-p_a)^3 + p(1-p_b)^3 \\
\mathbb{P}(Y=1) &= (1-p)3(1-p_a)^2 p_a + p3(1-p_b)^2 p_b \\
\mathbb{P}(Y=2) &= (1-p)3(1-p_a)p_a^2 + p3(1-p_b)p_b^2 \\
\mathbb{P}(Y=3) &= (1-p)p_a^3 + pp_b^3
\end{aligned}$$

With $\bar{\lambda}=(q,q_a,q_b)$, the identification system can be written:

$$F\left(\lambda,\bar{\lambda}\right)=0$$

$$\Longleftrightarrow$$

$$\begin{cases}
(1-p)(1-p_a)^3 + p(1-p_b)^3 = (1-q)(1-q_a)^3 + q(1-q_b)^3 \\
(1-p)3(1-p_a)^2 p_a + p3(1-p_b)^2 p_b = (1-q)3(1-q_a)^2 q_a + q3(1-q_b)^2 q_b \\
(1-p)3(1-p_a)p_a^2 + p3(1-p_b)p_b^2 = (1-q)3(1-q_a)q_a^2 + q3(1-q_b)q_b^2 \\
(1-p)p_a^3 + pp_b^3 = (1-q)q_a^3 + qq_b^3
\end{cases}$$

In order to illustrate the mechanism behind Theorem 1, we would like to compute the decomposition of $W=V(F)$, the zero-set of $F\left(\lambda,\bar{\lambda}\right)$ jointly in $\left(\lambda,\bar{\lambda}\right)$. While this geometric decomposition is hard to compute directly, we can rely on the strong correspondance between geometry and algebra that lies at the core of algebraic geometry, and carry an algebraic computation. $V(F)=V\left(\sqrt{\langle F\rangle}\right)$ where $\sqrt{\langle F\rangle}$ is the radical of the ideal generated by F, and the irreducible decomposition of W is 1-to-1 with the decomposition of $\sqrt{\langle F\rangle}$ in primes. We can compute the prime decomposition of

$\sqrt{\langle F \rangle}$ with Singular (Decker et al., 2015). It has 11 components:

$$\begin{aligned}
\sqrt{\langle F \rangle} = \langle p_b - q_b, p_a - q_a, p - q \rangle \cap \langle p_b - q_a, p_a - q_b, (1-p) - q \rangle & \quad I_1 \\
\cap \langle p_b - q_b, p_a - q_a, q_a - q_b \rangle & \quad I_2 \\
\cap \begin{cases} \langle p_b - q_b, q_a - q_b, p \rangle \cap \langle p_b - q_b, p_a - q_b, q \rangle \\ \cap \langle p_a - q_b, q_a - q_b, p - 1 \rangle \cap \langle p_b - q_a, p_a - q_a, q - 1 \rangle \end{cases} & \quad I_3 \\
\cap \begin{cases} \langle q - 1, p, p_b - q_a \rangle \cap \langle q, p, p_b - q_b \rangle \\ \cap \langle q - 1, p - 1, p_a - q_a \rangle \cap \langle q, p - 1, p_a - q_b \rangle \end{cases} & \quad I_4
\end{aligned}$$

Correspondingly:

$$W = \underbrace{V(I_1)}_{\text{generic region}} \cup \underbrace{V(I_2) \cup V(I_3) \cup V(I_4)}_{\text{exceptional region}} = V_g \cup V_e$$

$V(I_1)$ is the "nice" region, which contains the identified parameter values $\lambda = \bar{\lambda}$ as well as their label-switched versions. $V(I_2)$ contains parameter values for which coin flips happen with the same probability regardless of being in a/b or in $\lambda/\bar{\lambda}$. Of course, the first stage probabilities are not identified when this is the case. $V(I_3)$ contains parameter values for which in one $\lambda/\bar{\lambda}$ world, one of the flips is never observed, and in the other world, both flips happen but are indistinguishable due to having the same flipping probability. There are four subcomponents by symmetry and label-switching. $V(I_4)$ contains parameter values for which only one flip happens in both $\lambda/\bar{\lambda}$ worlds. There are also four subcomponents by symmetry and label-switching.

In this simple model, we can tell by eyeballing that $V_g = V(I_1)$ makes up the generic component of W, while $V_e = V(I_2) \cup V(I_3) \cup V(I_4)$ makes up its exceptional component. The exceptional region in $\bar{\lambda}$ space is simply the intersection of $\bar{\mathbb{A}}$ with the singular points V_e of W. The geometric object $V_e \cap \bar{\mathbb{A}}$ is not necessarily a closed set, but once more, we can use algebraic methods to compute its closure. We find:

$$\Lambda_e = \overline{V_e \cap \bar{\mathbb{A}}} = V(< (q_b - q_a)q(1-q) >)$$

2 Appendix for section 4: Asymptotics

2.1 Outline of the proofs

First, I prove the standard limit theorems (uniform law of large numbers for the log-likelihood, central limit theorem for the score and uniform law of large numbers for the observed information) assuming the data is stationary but allowing for a slightly misspecified likelihood computed with a wrong initial distribution μ, instead of the stationary $\mu^\diamond$. I use the "infinite-past" approach originally used in Baum and Petrie (1966) (see also Douc et al. (2014) for a textbook exposition and Douc et al. (2004) for a proof in the context of autoregressive hidden Markov models). A high-level description of the infinite-past strategy for the log-likelihood is as follows:

1. Write $L_T(\theta)$ as the sum of an auxiliary processes U_{1t}:

$$L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T-1} U_{1t}(\theta) \qquad \text{where } U_{1t}(\theta) = \log P_\theta(Y_{t+1}|Y_{1:t})$$

2. Show that the auxiliary process $U_{mt}(\theta) = \log P_\theta(Y_{t+1}|Y_{m:t})$ converges ($\theta^\star$ almost surely) to an "infinite-past" limit $U_t(\theta) = U_{-\infty t}(\theta)$ as $m \to -\infty$. $U_t(\theta)$ can to some extent be thought of as $\log P_\theta(Y_{t+1}|Y_{-\infty:t})$, although strictly speaking for $\theta \neq \theta^\star$ $P_\theta(Y_{t+1}|Y_{-\infty:t})$ is not well-defined in the usual sense because P_θ and $P_{\theta^\star}$ are mutually singular on the sequence space. The infinite-past limit is obtained by showing that $U_{mt}(\theta)$ is $\theta^\star$ almost surely Cauchy, uniformly in θ. The Cauchy bounds follow from non-homogeneous merging of the conditional Markov chain $(X_t|y_{m:T})_{t\geq m}$. Showing non-homogeneous merging is a key step and it is where the uniform lower bound $\underline{q}$ on the unobserved state transition matrix Q (assumption (A4)) is used.

3. Show that $\left\|L_T(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta)\right\| \xrightarrow{\theta^\star\ as} 0$. This follows directly from some bounds derived to show the Cauchy property.

4. Show a uniform law of large numbers for U_t, which is now an ergodic sum:

$$\left\|\frac{1}{T}\sum_{t=1}^{T-1} U_t(\theta) - L(\theta)\right\| \xrightarrow{\theta^\star\ as} 0$$

   For this I use a vector ergodic theorem valid for random vectors taking value

in separable metric spaces from Krengel (1985), sidestepping the need for an explicit argument around stochastic equicontinuity. The uniform law of large numbers is a pointwise law of large number in function space. When available, this argument may be of more general interest.

The same general infinite-past strategy is used for the score $s_T = \nabla_{\theta^\star} L_T(\theta)$ and the observed information $h_T = \nabla^2_\theta L_T(\theta)$, although new complications arise, such as making sure that some infinite sums are almost surely summable as we take m to $-\infty$. In order for the infinite-past to make sense, stationarity is key in this first step.

Second, I extend these limit theorems to the case where the data is nonstationary. I use an argument based on the merging properties of the chain. As far as I know, this argument is new and may be of more general interest.

Third, uniform local asymptotic normality of the model (Theorem 3) and asymptotic behavior of the maximum likelihood estimator (Theorem 4) follow in a standard fashion.

Fourth and finally, I prove the Bernstein–von Mises theorem for hidden Rust models Theorem 5 by checking that the assumptions of the general weakly dependent Bernstein–von Mises theorem from Connault (2014) are verified. One major motivation behind Connault (2014) was to obtain a Bernstein–von Mises theorem for regular time-series models in the spirit of Le Cam's (1986) theorem for smooth independent and identically distributed models, where local asymptotic normality is the main assumption. The existence of a uniformly consistent estimator is the second main assumption after local asymptotic normality and I appeal to the stable identification theorem (Theorem 2) to exhibit such a uniformly consistent estimator in the context of hidden Rust models.

2.2 Preliminary definitions and lemma about merging Markov chains

2.2.1 Merging Markov chains, concentration bounds

This section defines the *merging time* of a finite-state Markov chain, and states concentration inequalities for hidden Markov models from Paulin (2014). All definitions

are standard, although the terminology might vary. All results are known. General facts about finite-state Markov chains, including merging properties, can be found in Seneta (2006).

A Markov chain whose transition matrix changes with time is called a non-homogeneous Markov chain. A not necessarily homogeneous chain Z_t is *merging* when the total variation distance between the distributions of two independent chains started at arbitrary points goes to zero with time — the chains "merge":

$$\forall s, z_s, z'_s, \qquad d_{TV}(\mathcal{L}(Z_t|Z_s = z_s), \mathcal{L}(Z_t|Z_s = z'_s)) \xrightarrow[t\to\infty]{} 0$$

The following lemma (see, e.g., theorem 4.9 p.141 in Seneta (2006)) gives a sufficient condition for merging of a non-necessarily homogeneous Markov chain:

Lemma 1: Merging for (non-homogeneous) Markov chains under minorization
If the transition probabilities of a (not necessarily homogeneous) Markov chain Z_t are uniformly minorized in the sense that there are probability distributions ν_t and a constant $\underline{a} > 0$ such that for any values of the chain z_t and z_{t+1}:

$$P(z_{t+1}|z_t) \geq \underline{a}\nu_t(z_t + 1)$$

Then Z_t satisfies merging: for any two initial distributions μ_1 and μ_2:

$$d_{TV}\left(\mathcal{L}(Z_t|Z_0 \sim \mu_1), \mathcal{L}(Z_t|Z_0 \sim \mu_2)\right) < (1 - \underline{a})^t$$

If Z_t is homogeneous and merging, there is a distribution $\mu^\diamond$, necessarily unique, such that:

$$\forall z_1, \qquad d_{TV}(\mathcal{L}(Z_t|Z_1 = z_1), \mu^\diamond) \xrightarrow[t\to\infty]{} 0$$

Furthermore $\mu^\diamond$ is the unique stationary distribution of Z_t and the convergence happens geometrically fast: there is $\rho < 1$ and $c > 0$ such that:

$$\forall z_1, \qquad d_{TV}(\mathcal{L}(Z_t|Z_1 = z_1), \mu^\diamond) \leq c\rho^t$$

If Z_t is homogeneous and merging, under the minor additional assumption that Z_t has no transient states, or by ignoring the transient states (which correspond to zeros in $\mu^\diamond$), Z_t is irreducible aperiodic. Conversely, an irreducible aperiodic chain is recurrent

and merging. The merging time is a quantity used to measure the merging speed. The ϵ-merging time $\tau_z(\epsilon)$ of a merging Markov chain z is defined as follows, for $0 < \epsilon < 1$:

$$\tau_z(\epsilon) = \min\{t : \max_z\{d_{TV}(\mathcal{L}(Z_t|Z_1 = z_1), \mu^\diamond)\} < \epsilon\}$$

The absolute merging time, or simply *merging time*, is:

$$\tau_z = \inf_{0\leq\epsilon<1} \frac{\tau_z(\epsilon/2)}{(1-\epsilon)^2}$$

This seemingly ad-hoc definition is the convenient one for stating concentration inequalities where the concentration constant is directly proportional to the merging time. I state McDiarmid inequalities for (not necessarily stationary) hidden Markov models from Paulin (2014) (corollary 2.14 p.12):

Lemma 2: Concentration inequalities for HMMs
For (x, y) a HMM, write τ_x for the merging time of the Markov chain x. Let f be a function of T arguments with bounded differences:

$$f(y_1, ..., y_T) - f(y'_1, ..., y'_T) \leq \sum_{t=1}^{T} c_t 1[y_t \neq y'_t]$$

Then the following one-sided and two-sided concentration inequalities hold:

$$P\left(f < \mathbb{E}\left[f\right] - u\right) \leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_x \sum_{t=1}^{T} c_t^2}\right) \tag{1}$$

$$P\left(f > \mathbb{E}\left[f\right] + u\right) \leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_x \sum_{t=1}^{T} c_t^2}\right) \tag{2}$$

$$P\left(|f - \mathbb{E}[f]| > u\right) \leq 2\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x \sum_{t=1}^{T} c_t^2}\right) \tag{3}$$

According to 1, concentration is "$4\tau_x$ times slower" for a merging Markov chain than for a sequence of independent and identically distributed random variables.

2.2.2 z is uniformly merging

In this section I prove that z is uniformly merging, meaning that the merging time of $z_t = (x_t, s_t, a_t)$ is uniformly bounded over the parameter space.

We know that Θ is compact and that for any θ, z is recurrent and merging. We want to show that z is *uniformly merging*, in the sense that the merging time is uniformly bounded: there is $\bar{\tau}_z < \infty$ such that:

$$\forall \theta \in \Theta, \qquad \tau_z(\theta) < \bar{\tau}_z$$

It is not clear that τ_z itself is a continuous function of the transition matrix but we will use a bound from Paulin (2014).

Because z is merging, z has a unique marginal stationary distribution, which we write $\mu^\diamond(\theta)$. Because z is recurrent, $\mu^\diamond(\theta) > 0$.

Define $\lambda(\theta)$ as the second biggest eigenvalue of the multiplicative reversibilization $\tilde{P}_z$ of the transition matrix P_z of z (see Fill (1991)). $\tilde{P}_z$ is a continuous function of P_z, and $\tilde{P}_z$ is recurrent merging because P_z is.

A consequence of bound (3.12) p. 16 together with bound (2.9) p. 10 in Paulin (2014) is:

$$\tau_z(\theta) \leq 4 \frac{1 + 2\log 2 + \log 1/\mu^\diamond_{min}(\theta)}{1 - \lambda(\theta)}$$

Stationary distributions and eigenvalues are continuous functions of matrix coefficients. As a consequence, $\lambda(\theta)$ is bounded away from 1 and $\mu^\diamond(\theta)$ is bounded away from zero on Θ. The conclusion follows. We call τ_z the lowest uniform upper bound:

$$\tau_z = \sup_{\theta \in \Theta} \tau_z(\theta) < \infty$$

For the same reasons, blocks of z's are also uniformly merging. Let R be any natural number ≥ 1 and $\hat{z}_s$ be non-overlapping consecutive blocks of R z_t's: $\hat{z}_1 = (z_1, \ldots, z_R), \hat{z}_2 = (z_{R+1}, \ldots, z_{2R})$, etc. $\hat{z}$ is merging because z is merging and consequently $\hat{z}$ is uniformly merging for the same reason z is uniformly merging. We write $\tau_{\hat{z}}$ for the corresponding uniform merging time ($\tau_{\hat{z}}$ can depend on R).

2.3 Limit theorems for stationary hidden Rust models

In this section, assume (X_t, Y_t) are distributed according to the stationary distribution induced by $\theta^\star$ (recall that, by the merging assumption (A2), each θ induces a unique marginal stationary distribution $\mu^\diamond(\theta)$). Under this assumption, (X_t, Y_t) can be extended to $(X, Y)_{-\infty:+\infty}$.

The econometrician maximizes a conditional log-likelihood potentially misspecified in terms of the initial distribution:

$$L_T(\theta) = \frac{1}{T} \log P_{\theta,\mu} \left(Y_{2:T} | Y_1 \right)$$

$P_{\theta,\mu}$ means the value of the probability under transitions indexed by θ and initial distribution μ (on the earliest index appearing in $P_{\theta,\mu}(\cdot)$). In fact, μ plays no role in the proof and is suppressed from notation for simplicity.

2.3.1 Uniform law of large numbers for the log-likelihood

This section shows a uniform law of large numbers for the log-likelihood:

$$\| L_T(\theta) - L(\theta) \| \xrightarrow{\theta^\star \ as} 0$$

where $L(\theta)$ can be thought of as $\mathbb{E}_{\theta^\star}[\log P_\theta \left(Y_1 | Y_{-\infty:0} \right)]$. The following decomposition will be used:

$$\begin{cases} L_T(\theta) = \frac{1}{T} \sum_{t=1}^{T-1} U_{1t}(\theta) \\ \left\| \frac{1}{T} \sum_{t=1}^{T-1} U_{1t}(\theta) - \frac{1}{T} \sum_{t=1}^{T-1} U_t(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T} \sum_{t=1}^{T-1} U_t(\theta) - L(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \end{cases}$$

Where $U_t(\theta)$ will be defined as the limit of $U_{mt}(\theta) := \log P_\theta \left(Y_{t+1} | Y_{m:t} \right)$ as $m \to -\infty$. $U_t(\theta)$ will be shown to be stationary ergodic and we will have $L(\theta) = \mathbb{E}_{\theta^\star}[U_0(\theta)]$. There are three steps in the proof:

1. $U_{mt}(\theta)$ is an (almost surely) Cauchy sequence in $m \to -\infty$, uniformly in θ. Call $U_t(\theta)$ its limit.
2. $\left\| \frac{1}{T} \sum_{t=1}^{T-1} U_{1t}(\theta) - \frac{1}{T} \sum_{t=1}^{T-1} U_t(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$.
3. $\left\| \frac{1}{T} \sum_{t=1}^{T-1} U_t(\theta) - L(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$.

Step 1: U_{mt} is $\theta^\star$ almost surely uniform Cauchy

We want to define $U_t(\theta) = U_{-\infty t}(\theta)$ and show the following $\theta^\star$ almost sure geometric bound, valid for any $-\infty \leq m < m' \leq 1$ and uniform in θ:

$$|U_{m't}(\theta) - U_{mt}(\theta)| \leq K\rho^{t-m'} \tag{4}$$

First, we show (4) for $-\infty < m < m' \leq 1$. Note that because $|\log x - \log y| \leq \frac{|x-y|}{x \wedge y}$:

$$|\log P_\theta(y_{t+1}|y_{m':t}) - \log P_\theta(y_{t+1}|y_{m:t})| \leq \frac{|P_\theta(y_{t+1}|y_{m':t}) - P_\theta(y_{t+1}|y_{m:t})|}{P_\theta(y_{t+1}|y_{m':t}) \wedge P_\theta(y_{t+1}|y_{m:t})}$$

A lower bound for the denominator is easy to find. Note that:

$$P_\theta(y_{t+1}|y_{m:t}) = \sum_{x_{t+1},x_t} P_\theta(y_{t+1}|x_{t+1},y_t)P_\theta(x_{t+1}|x_t)P_\theta(x_t|y_{m:t})$$

So that:

$$P_\theta(y_{t+1}|y_{m:t}) \geq \underline{q} \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t)$$

Let us turn to the numerator. By conditional independence, $(X_t|y_{m:T})_{t\geq m}$ is (non-homogeneous) Markov. We show that it satisfies a merging property uniformly in θ:

Lemma 3: Uniform merging for $(X_t|y_{m:T})_{t\geq m}$
There is $\rho < 1$ such that for any m, for any two initial distributions μ_1 and μ_2 on X_m, for any θ, the following inequality holds (for any $y_{m:T}$ with positive probability):

$$d_{TV}\left(\mathcal{L}_\theta(X_t|y_{m:T};\mu_1), \mathcal{L}_\theta(X_t|y_{m:T};\mu_2)\right) < \rho^{t-m}$$

Proof. For any $t \geq m$, any $(x_t, y_{m:T})$ with positive probability:

$$\begin{aligned} P_\theta(x_{t+1}|x_t, y_{m:T})P_\theta(y_{t+1:T}|x_t, y_{m:t}) &= P_\theta(x_{t+1}, y_{t+1:T}|x_t, y_{m:t}) \\ &= P_\theta(y_{t+1:T}|x_{t+1}, x_t, y_{m:t})P_\theta(x_{t+1}|x_t, y_{m:t}) \\ &= P_\theta(y_{t+1:T}|x_{t+1}, y_t)P_\theta(x_{t+1}|x_t) \\ P_\theta(x_{t+1}|x_t, y_{m:T}) &= \frac{P_\theta(y_{t+1:T}|x_{t+1}, y_t)P_\theta(x_{t+1}|x_t)}{\sum_{x_{t+1}} P_\theta(y_{t+1:T}|x_{t+1}, y_t)P_\theta(x_{t+1}|x_t)} \\ &\geq \underline{q}\frac{P_\theta(y_{t+1:T}|x_{t+1}, y_t)}{\sum_{x_{t+1}} P_\theta(y_{t+1:T}|x_{t+1}, y_t)} \end{aligned}$$

$\nu_t(x_{t+1}; \theta) := \frac{P_\theta(y_{t+1:T}|x_{t+1}, y_t)}{\sum_{x_{t+1}} P_\theta(y_{t+1:T}|x_{t+1})}$ defines a probability distribution on X_{t+1}, and the above inequality is a uniform minorization of the transition probabilities by $\nu_t(\theta)$:

$$P_\theta(x_{t+1}|x_t, y_{m:T}) \geq \underline{q}\nu_t(x_{t+1}; \theta)$$

By Lemma 1 in section 2.2:

$$d_{TV}\left(\mathcal{L}_\theta(X_t|y_{m:T}; \mu_1), \mathcal{L}_\theta(X_t|y_{m:T}; \mu_2)\right) < (1 - \underline{q})^{t-m}$$

□

Coming back to bounding the numerator, remember that for any two probabilities μ_1 and μ_2 and any f, $0 \leq f \leq 1$:

$$|\mu_1 f - \mu_2 f| \leq d_{TV}\left(\mu_1, \mu_2\right) \tag{5}$$

Then, for any $y_{m:T}$ with positive probability:

$$
\begin{aligned}
&|P_\theta(y_{t+1}|y_{m':t}) - P_\theta(y_{t+1}|y_{m:t})| \\
&\quad = \left| \sum_{x_{t+1},x_t} P_\theta(y_{t+1}|x_{t+1},y_t) P_\theta(x_{t+1}|x_t)(P_\theta(x_t|y_{m':t}) - P_\theta(x_t|y_{m:t})) \right| \\
&\quad \leq \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t) \left| \sum_{x_t} P_\theta(x_{t+1}|x_t)(P_\theta(x_t|y_{m':t}) - P_\theta(x_t|y_{m:t})) \right| \\
&\quad \leq \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t) \left| \sum_{x_t} P_\theta(x_{t+1}|x_t) \left(P_\theta(x_t|y_{m':t}) - \sum_{x_{m'}} P_\theta(x_t|y_{m':t},x_{m'}) P_\theta(x_{m'}|y_{m:t}) \right) \right| \\
&\quad \leq \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t) d_{TV}\left(\mathcal{L}_\theta(X_t|y_{m':t};x_{m'}|y_{m':t}), \mathcal{L}_\theta(X_t|y_{m':t};x_{m'}|y_{m:t})\right) \quad \text{by (5)} \\
&\quad \leq \rho^{t-m'} \sum_{x_{t+1}} P_\theta(y_{t+1}|x_{t+1},y_t) \qquad \text{by merging, lemma 3}
\end{aligned}
$$

Putting bounds for the numerator and denominator together, for $-\infty < m < m' \leq 1$ we have (almost surely):

$$
|\log P_\theta\left(Y_{t+1}|Y_{m':t}\right) - \log P_\theta\left(Y_{t+1}|Y_{m:t}\right)| \leq \frac{\rho^{t-1} \sum_{x_{t+1}} P_\theta\left(Y_{t+1}|X_{t+1},Y_t\right)}{\underline{q} \sum_{x_{t+1}} P_\theta\left(Y_{t+1}|X_{t+1},Y_t\right)} = \frac{\rho^{t-m'}}{\underline{q}} \quad (6)
$$

Note that we can collect all the (countable) null sets of (6) so that the precise statement is that the inequality holds "almost surely: for all m and m'" and not "for each m and m': almost surely." This implies that $U_{mt}(\theta)$ is almost surely a Cauchy sequence as $m \to -\infty$. As a consequence there is U_t such that (almost surely):

$$
U_{mt}(\theta) \xrightarrow[m\to-\infty]{} U_t(\theta)
$$

By continuity, (4) holds for $m = -\infty$ too.

Step 2: $\left\| \frac{1}{T} \sum_{t=1}^{T-1} U_{1t}(\theta) - \frac{1}{T} \sum_{t=1}^{T-1} U_t(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$

From the (almost sure) geometric bound (4) (with $m = -\infty$ and $m' = 1$), we have (almost surely):

$$\left|\frac{1}{T}\sum_{t=1}^{T-1}\log P_\theta\left(Y_{t+1}|Y_{1:t}\right)-\frac{1}{T}\sum_{t=1}^{T-1}\log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right)\right|$$
$$\leq \frac{1}{T}\sum_{t=1}^{T-1}\left|\log P_\theta\left(Y_{t+1}|Y_{1:t}\right)-\log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right)\right|$$
$$\leq \frac{1}{T}\frac{1}{\underline{q}}\sum_{t=1}^{T-1}\rho^{t-1}$$
$$\leq \frac{1}{T}\frac{1}{\underline{q}}\frac{1}{1-\rho}$$

which implies:

$$\left\|\frac{1}{T}\sum_{t=1}^{T-1}U_{1t}(\theta)-\frac{1}{T}\sum_{t=1}^{T-1}U_t(\theta)\right\|\xrightarrow{\theta^\star\ as}0$$

Step 3: $\left\|\frac{1}{T}\sum_{t=1}^{T-1}U_t(\theta)-L(\theta)\right\|\xrightarrow{\theta^\star\ as}0$

This is a consequence of the almost sure ergodic theorem in function space. To see this, note that $\theta\to U_{mt}(\theta)$ is (almost surely) continuous because $P_\theta\left(y_{t+1}|y_{m:t}\right)$ is a rational function of the transition matrices' coefficients. The results of the first step imply that $\theta\to U_t(\theta)$ is (almost surely) continuous as a uniform Cauchy limit. Thus we can consider all of them to be everywhere continuous without loss of generality. Now define $\mathcal{C}(\Theta)$ to be the set of continuous functions from Θ to $\mathbb{R}$ and:

$$\begin{array}{lcll} s:\mathcal{Y}^{\mathbb{Z}} & \longrightarrow & \mathcal{Y}^{\mathbb{Z}} & \text{(the shift operator)} \\ (y_t)_{t\in\mathbb{Z}} & \longrightarrow & (y_{t+1})_{t\in\mathbb{Z}} & \\ l:\mathcal{Y}^{\mathbb{Z}} & \longrightarrow & \mathcal{C}(\Theta) & \\ (y_t)_{t\in\mathbb{Z}} & \longrightarrow & U_0(\theta)=P_\theta(y_1|y_{-\infty:0}) & \end{array}$$

Using the standard notation:

$$s^t l = l\circ\underbrace{s\circ\ldots\circ s}_{t\text{ times}}$$

we can rewrite:

$$\frac{1}{T}\sum_{t=1}^{T-1}\log P_\theta\left(Y_{t+1}|Y_{-\infty:t}\right)=\frac{1}{T}\sum_{t=1}^{T-1}s^t l\left(Y\right)$$

Y is stationary ergodic: this is exactly the setting of the ergodic theorem. In the most familiar case, l would be a measurable function from $\mathcal{Y}^{\mathbb{Z}}$ to $\mathbb{R}$, which is not the case here. However, $(\mathcal{C}(\Theta), \|\cdot\|_\infty)$ is separable because Θ is compact: a *vector* almost sure ergodic theorem holds, exactly similar to the scalar case (see theorem 2.1 p.167 in Krengel (1985)).

Thus:

$$\left\| \frac{1}{T} \sum_{t=1}^{T-1} U_t(\theta) - L(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$$

where:

$$L = \mathbb{E}_{\theta^\star} \left[l\left(Y\right) \right] = \mathbb{E}_{\theta^\star} \left[\log P_\theta \left(Y_1 | Y_{-\infty:0} \right) \right] \in \mathcal{C}(\Theta)$$

2.3.2 Central limit theorem for the score

Let $s_T = \nabla_{\theta^\star} L_T(\theta)$ be the (observed) score. This section shows a central limit theorem (pointwise at $\theta^\star$) for the score: there is I such that:

$$\sqrt{T} s_T \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I)$$

Similar to the proof of the uniform law of large numbers for the log-likelihood (section 2.3.1), the following decomposition is used:

$$\begin{cases} s_T = \frac{1}{T} \sum_{t=1}^{T-1} V_{1t} + o_{\theta^\star}\left(\frac{1}{\sqrt{T}}\right) \\ \left| \sqrt{T} \frac{1}{T} \sum_{t=1}^{T-1} V_{1t} - \sqrt{T} \frac{1}{T} \sum_{t=1}^{T-1} V_t \right| \xrightarrow{\theta^\star \ as} 0 \\ \sqrt{T} \frac{1}{T} \sum_{t=1}^{T-1} V_t \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I) \end{cases}$$

V_t is the "infinite-past" limit of an auxiliary process V_{mt} (to be defined shortly) as $m \to -\infty$.

In order to define the auxiliary process V_{mt}, note an identity of Louis (1982) (equation (3.1) p. 227), which says that in a model where Y is observed and X is unobserved:

$$\nabla_{\theta^\star} \log P_\theta(Y) = \mathbb{E}_{\theta^\star}[\nabla_{\theta^\star} \log P_\theta(Y, X) | Y] \tag{7}$$

Here:

$$
\begin{aligned}
s_T &= \frac{1}{T}\nabla_{\theta^\star}\log P_\theta\left(Y_{2:T}|Y_1\right) \\
&= \frac{1}{T}\mathbb{E}_{\theta^\star}\left[\nabla_{\theta^\star}\log P_\theta\left(Y_{2:T},X_{1:T}|Y_1\right)|Y_{1:T}\right] && \text{by (7)} \\
&= \frac{1}{T}\mathbb{E}_{\theta^\star}\left[\nabla_{\theta^\star}\sum_{s=1}^{T-1}\log P_\theta\left(X_{s+1},Y_{s+1}|X_s,Y_s\right)\middle|Y_{1:T}\right] \\
&\qquad + \underbrace{\frac{1}{T}\underbrace{\mathbb{E}_{\theta^\star}\left[\nabla_{\theta^\star}\log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right]}_{\xrightarrow{\theta^\star\ as}\mathbb{E}_{\theta^\star}[\cdot|Y_{1:+\infty}]}}_{=o_{\theta^\star}\left(\frac{1}{\sqrt{T}}\right)} && \text{by conditional independence}
\end{aligned}
$$

Write $J_s = \nabla_{\theta^\star}\log P_\theta\left(X_{s+1},Y_{s+1}|X_s,Y_s\right)$ and consider the telescopic sum:

$$
\begin{aligned}
\mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-1}J_s\middle|Y_{1:T}\right] = \quad & \mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-1}J_s\middle|Y_{1:T}\right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-2}J_s\middle|Y_{1:T-1}\right] && (=V_{1,T-1}) \\
& + \mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-2}J_s\middle|Y_{1:T-1}\right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=1}^{T-3}J_s\middle|Y_{1:T-2}\right] && (=V_{1,T-2}) \\
& + \dots \\
& + \mathbb{E}_{\theta^\star}\left[J_1|Y_{1:2}\right] && (=V_{1,1})
\end{aligned}
$$

Finally introduce the auxiliary process:

$$
\begin{aligned}
V_{mt} &= \mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t}J_s\middle|Y_{m:t+1}\right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t-1}J_s\middle|Y_{m:t}\right] \\
&= \mathbb{E}_{\theta^\star}\left[J_t|Y_{m:t+1}\right] + \sum_{s=m}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t+1}\right] - \mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t}\right]\right)
\end{aligned}
$$

As announced:

$$
s_T = \frac{1}{T}\sum_{t=1}^{T-1}V_{1t} + o_{\theta^\star}\left(\frac{1}{\sqrt{T}}\right)
$$

The remainder of the proof proceeds in three steps:

1. V_{mt} is a ($\theta^\star$ almost-sure) Cauchy sequence for $m \to -\infty$. We call V_t its limit.
2. $\left|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_{1t} - \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_t\right| \xrightarrow{\theta^\star\ as} 0$.
3. $\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_t \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0,I)$.

Step 1: V_{mt} is $\theta^\star$ almost surely Cauchy

Similar to step 1 in section 2.3.1, we want to define $V_t = V_{-\infty t}$ and show the following $\theta^\star$ almost sure inequality for $-\infty \leq m < m' \leq 1$:

$$|V_{m't} - V_{mt}| \leq K\rho^{t-m'} \tag{8}$$

For now fix $-\infty < m < m' \leq 1$. We split the sum $|V_{m't} - V_{mt}|$ into four regions. Call $k = \lfloor \frac{m'+t}{2} \rfloor$.

$$\begin{aligned}
|V_{m't} - V_{mt}| = &\left| \left(\mathbb{E}_{\theta^\star}\left[\sum_{s=m'}^{t} J_s \middle| Y_{m':t+1} \right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=m'}^{t-1} J_s \middle| Y_{m':t} \right] \right) \right. \\
&\left. - \left(\mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t} J_s \middle| Y_{m:t+1} \right] - \mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t-1} J_s \middle| Y_{m:t} \right] \right) \right| \\
\leq &\sum_{s=k}^{t} \left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t+1} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t+1} \right] \right| + \sum_{s=k}^{t-1} \left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t} \right] \right| \\
&+ \sum_{s=m'}^{k-1} \left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t+1} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t} \right] \right| + \sum_{s=m}^{k-1} \left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t+1} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t} \right] \right|
\end{aligned}$$

The geometric bound (8) for V_{mt} is then a consequence of the following $\theta^\star$ almost sure geometric bounds, one for each region:

$$\left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t+1} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t+1} \right] \right| \leq K\rho^{s-m'} \tag{9}$$

$$\left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t} \right] \right| \leq K\rho^{s-m'} \tag{10}$$

$$\left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t+1} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t} \right] \right| \leq K\rho^{t-s} \tag{11}$$

$$\left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t+1} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t} \right] \right| \leq K\rho^{t-s} \tag{12}$$

(9) is a consequence of the merging properties of $(X_t | y_{m:T})_{t \geq m}$, which were proven in lemma 3. Indeed $\mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t+1} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t+1} \right] = \mathbb{E}_{\theta^\star}\left[\mathbb{E}_{\theta^\star}[J_s | X_s, Y_s] | Y_{m':t+1} \right] - \mathbb{E}_{\theta^\star}\left[\mathbb{E}_{\theta^\star}[J_s | X_s, Y_s] | Y_{m:t+1} \right]$, and J_s is uniformly bounded by compactness of Θ and smoothness of the model, so that:

$$\left| \mathbb{E}_{\theta^\star}\left[J_s | Y_{m':t+1} \right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{m:t+1} \right] \right| \leq K d_{TV}\left(\mathcal{L}_\theta(X_s | y_{m':t}; x_{m'} | y_{m':t}), \mathcal{L}_\theta(X_t | y_{m':t}; x_{m'} | y_{m:t}) \right)$$

Similarly (10) is also a consequence of the merging properties of $(X_t | y_{m:T})_{t \geq m}$. Bounds

(11) and (12) are a consequence of the merging properties of another Markov chain, namely $(X_{T-t}|y_{m:T})_{0\leq t\leq T-m}$ (note the reverse time). The merging of $(X_{T-t}|y_{m:T})_{0\leq t\leq T-m}$ as well as the bounds (9), (10), (11) and (12), is proven similarly to the corresponding results in the log-likelihood section. The proofs are omitted for brevity.

Note that (9), (10) and (12) extend to $m=-\infty$ because:

$$\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t}\right] \xrightarrow{\theta^\star\ as} \mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]$$

As a consequence, $\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]$ is (almost surely) absolutely summable in $s\to-\infty$ and $\sum_{s=-\infty}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]\right)$ is well-defined. We can legitimately define:

$$V_t = V_{-\infty t} = \mathbb{E}_{\theta^\star}\left[J_t|Y_{-\infty:t+1}\right] + \sum_{s=-\infty}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{-\infty:t}\right]\right)$$

(In a few lines it will be shown that $V_t=V_{-\infty t}$ is the (almost sure) limit of V_{mt} as $m\to-\infty$.)

With this definition, the inequality:

$$\begin{aligned}|V_{m't}-V_{mt}| \leq \sum_{s=k}^{t}\left|\mathbb{E}_{\theta^\star}\left[J_s|Y_{m':t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t+1}\right]\right| + \sum_{s=k}^{t-1}\left|\mathbb{E}_{\theta^\star}\left[J_s|Y_{m':t}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t}\right]\right| \\ + \sum_{s=m'}^{k-1}\left|\mathbb{E}_{\theta^\star}\left[J_s|Y_{m':t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{m':t}\right]\right| + \sum_{s=m}^{k-1}\left|\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t+1}\right]-\mathbb{E}_{\theta^\star}\left[J_s|Y_{m:t}\right]\right|\end{aligned}$$

is valid for $m=-\infty$ too, and by (9), (10), (11) and (12):

$$\begin{aligned}|V_{m't}-V_{mt}| &\leq \sum_{s=k}^{t} c\rho^{s-m'} + \sum_{s=k}^{t-1} c\rho^{s-m'} + \sum_{s=m'}^{k-1} c\rho^{t-s} + \sum_{s=m}^{k-1} c\rho^{t-s} \\ &\leq c(\rho^{k-m'}+\rho^{k-m'}+\rho^{t-k}+\rho^{t-k}) \\ &\leq c\sqrt{\rho}^{t-m'}\end{aligned}$$

In particular, $V_{mt}\to V_t$ as $m\to-\infty$.

Step 2: $\left|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_{1t} - \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}V_t\right| \xrightarrow{\theta^\star\ as} 0$

From the geometric bound (8) (with $m = -\infty$ and $m' = 1$), we have (almost surely):

$$\begin{aligned}\left|\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_{1t} - \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_t\right| &\leq \sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}|V_{1t} - V_t| \\ &\leq K\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1}\rho^{t-1} \\ &\leq K\sqrt{T}\frac{1}{T}\frac{1}{1-\rho} \\ &\xrightarrow[T\to\infty]{} 0\end{aligned}$$

Step 3: $\frac{1}{T}\sum_{t=1}^{T-1} V_t \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I)$
Remember that:

$$V_t = \mathbb{E}_{\theta^\star}\left[J_t | Y_{-\infty:t+1}\right] + \sum_{s=-\infty}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s | Y_{-\infty:t+1}\right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{-\infty:t}\right]\right)$$

V_t is immediately ergodic stationary because (X, Y) is. We show that it is also a martingale difference sequence with respect to the $\sigma\left(Y_{-\infty:t+1}\right)$ filtration. Note that for $m > -\infty$:

$$\begin{aligned}\mathbb{E}_{\theta^\star}\left[\sum_{s=m}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s | Y_{-\infty:t+1}\right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{-\infty:t}\right]\right)\middle| Y_{-\infty:t}\right] &= \sum_{s=m}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s | Y_{-\infty:t}\right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{-\infty:t}\right]\right) \\ &= 0\end{aligned}$$

Thus by dominated convergence:

$$\mathbb{E}_{\theta^\star}\left[\sum_{s=-\infty}^{t-1}\left(\mathbb{E}_{\theta^\star}\left[J_s | Y_{-\infty:t+1}\right] - \mathbb{E}_{\theta^\star}\left[J_s | Y_{-\infty:t}\right]\right)\middle| Y_{-\infty:t}\right] = 0$$

So that:

$$\begin{aligned}\mathbb{E}_{\theta^\star}\left[V_t | Y_{-\infty:t}\right] &= \mathbb{E}_{\theta^\star}\left[J_t | Y_{-\infty:t}\right] + 0 \\ &= \mathbb{E}_{\theta^\star}\left[\mathbb{E}_{\theta^\star}\left[d\log P\left(X_{t+1}, Y_{t+1} | X_t, Y_t\right) | X_t, Y_t\right] | Y_{-\infty:t}\right] \\ &= 0 \qquad \text{(expectation of the conditional score)}\end{aligned}$$

By the central limit theorem for ergodic stationary martingale difference sequences:

$$\sqrt{T}\frac{1}{T}\sum_{t=1}^{T-1} V_t \overset{\theta^\star}{\rightsquigarrow} \mathcal{N}(0, I) \qquad \text{where } I = \mathbb{E}_{\theta^\star}[V_1 V_1']$$

2.3.3 Uniform law of large numbers for the observed information

Let $i_T(\theta) = -\nabla^2_\theta L_T(\theta)$ be the (observed) information. This section shows a uniform law of large numbers for the information:

$$\|i_T(\theta) - i(\theta)\| \xrightarrow{\theta^\star \ as} 0$$

where $i(\theta)$ will be defined below and $i(\theta^\star) = I$ is the asymptotic variance of the score (see section 2.3.2).

Similar to the proofs of the uniform law of large numbers for the log-likelihood (section 2.3.1) and of the central limit theorem for the score (section 2.3.2), the following decomposition is used:

$$\begin{cases} i_T(\theta) = \frac{1}{T}\sum_{t=1}^{T-1} W^e_{1t}(\theta) + \frac{1}{T}\sum_{t=1}^{T-1} W^v_{1t}(\theta) + \frac{1}{T}\sum_{t=1}^{T-1} W^c_{1t}(\theta) + o_{\theta^\star}(1) \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^e_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^e_{t}(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^v_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^v_{t}(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^c_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^c_{t}(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^e_{t}(\theta) - i^e(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^v_{t}(\theta) - i^v(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \\ \left\| \frac{1}{T}\sum_{t=1}^{T-1} W^c_{t}(\theta) - i^c(\theta) \right\| \xrightarrow{\theta^\star \ as} 0 \end{cases}$$

$W^e_t(\theta)$, $W^v_t(\theta)$ and $W^c_t(\theta)$ are the uniform Cauchy "infinite-past" limits of the auxiliary processes $W^e_{mt}(\theta)$, $W^v_{mt}(\theta)$ and $W^c_{mt}(\theta)$ (to be defined shortly) respectively, as $m \to -\infty$.

In order to define the auxiliary processes $W^e_{mt}(\theta)$, $W^v_{mt}(\theta)$ and $W^c_{mt}(\theta)$, note another identity of Louis (1982) (equation (3.2) p. 227), which says that in a model where Y

is observed and X is unobserved::

$$\nabla^2_{\theta^\star} \log P_\theta(Y) = \mathbb{E}_{\theta^\star}[\nabla^2_{\theta^\star} \log P_\theta(Y,X)|Y] + V_{\theta^\star}[\nabla_{\theta^\star} \log P_\theta(Y,X)|Y] \tag{13}$$

Define $J_s(\theta) = \nabla_\theta \log P_\theta\left(X_{s+1}, Y_{s+1}|X_s, Y_s\right)$ and $H_s(\theta) = \nabla^2_\theta \log P_\theta\left(X_{s+1}, Y_{s+1}|X_s, Y_s\right)$. In particular, J_s in the previous section is $J_s(\theta^\star)$ according to this definition. In the following $E_\theta[Z_1|Z_2]$ means a version of Kolmogorov's conditional expectation under θ evaluated at Z_2, but note that Z_2 will typically be distributed according to $\theta^\star \neq \theta$.

Now by (13) and conditional independence:

$$\begin{aligned}
i_T(\theta) &= \frac{1}{T}\nabla^2_\theta \log P_\theta\left(Y_{2:T}|Y_1\right) \\
&= \frac{1}{T}E_\theta\left[\nabla^2_\theta \log P_\theta\left(Y_{2:T}, X_{1:T}|Y_1\right)\middle|Y_{1:T}\right] + \frac{1}{T}V_\theta\left[\nabla_\theta \log P_\theta\left(Y_{2:T}, X_{1:T}|Y_1\right)|Y_{1:T}\right] \\
&= \frac{1}{T}E_\theta\left[\sum_{s=1}^{T-1} H_s(\theta)\middle|Y_{1:T}\right] + \frac{1}{T}V_\theta\left[\sum_{s=1}^{T-1} J_s(\theta)\middle|Y_{1:T}\right] + \frac{1}{T}E_\theta\left[\nabla^2_\theta \log P_\theta\left(X_1|Y_1\right)\middle|Y_{1:T}\right] \\
&\qquad + \frac{1}{T}2Cov_\theta\left[\sum_{s=1}^{T-1} J_s(\theta), \nabla_\theta \log P_\theta\left(X_1|Y_1\right)\middle|Y_{1:T}\right] + \frac{1}{T}V_\theta\left[\nabla_\theta \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right]
\end{aligned}$$

$E_\theta\left[\nabla^2_\theta \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right]$ and $V_\theta\left[\nabla_\theta \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right]$ converge to some "infinite-past" limit by Cauchy-ness, using the same proof strategy we have used repeatedly, so that:

$$\frac{1}{T}E_\theta\left[\nabla^2_\theta \log P_\theta\left(X_1|Y_1\right)\middle|Y_{1:T}\right] + \frac{1}{T}V_\theta\left[\nabla_\theta \log P_\theta\left(X_1|Y_1\right)|Y_{1:T}\right] = o_{\theta^\star}(1)$$

We introduce the auxiliary processes:

$$\begin{aligned}
W^e_{mt}(\theta) &= E_\theta\left[\sum_{s=m}^{t} H_s(\theta)\middle|Y_{m:t+1}\right] - E_\theta\left[\sum_{s=m}^{t-1} H_s(\theta)\middle|Y_{m:t}\right] \\
W^v_{mt}(\theta) &= V_\theta\left[\sum_{s=m}^{t} J_s(\theta)\middle|Y_{m:t+1}\right] - V_\theta\left[\sum_{s=m}^{t-1} J_s(\theta)\middle|Y_{m:t}\right] \\
W^c_{mt}(\theta) &= Cov_\theta\left[\sum_{s=m}^{t} J_s(\theta), \nabla_\theta \log P_\theta\left(X_1|Y_1\right)\middle|Y_{m:t+1}\right] - Cov_\theta\left[\sum_{s=m}^{t-1} J_s(\theta), \nabla_\theta \log P_\theta\left(X_1|Y_1\right)\middle|Y_{m:t}\right]
\end{aligned}$$

Using telescopic sums similar to what was done for the score in section 2.3.2, we have,

as announced:

$$i_T(\theta) = \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^e(\theta) + \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^v(\theta) + \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^c(\theta) + o_{\theta^\star}(1)$$

The remainder of the proof proceeds in three steps:

1. W_{mt}^e, W_{mt}^v and W_{mt}^c are $\theta^\star$ almost sure uniform Cauchy sequences as $m \to -\infty$. We call $W_t^e(\theta)$, $W_t^v(\theta)$ and $W_t^c(\theta)$ their limits.
2. $\left\| \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^e(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_t^e(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$, $\left\| \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^v(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_t^v(\theta) \right\| \xrightarrow{\theta^\star \ as}$ 0 and $\left\| \frac{1}{T}\sum_{t=1}^{T-1} W_{1t}^c(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W_t^c(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$.
3. $\left\| \frac{1}{T}\sum_{t=1}^{T-1} W_t^e(\theta) - i^e(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$, $\left\| \frac{1}{T}\sum_{t=1}^{T-1} W_t^v(\theta) - i^v(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$ and $\left\| \frac{1}{T}\sum_{t=1}^{T-1} W_t^c(\theta) - i^c(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$.

Step 1.1: $W_{mt}^e(\theta)$ is $\theta^\star$ almost surely uniform Cauchy

Similar to step 1 in the log-likelihood section (section 2.3.1) and step 1 in the score section (section 2.3.2), we define $W_t^e(\theta) = W_{-\infty t}^e(\theta)$ and show the following $\theta^\star$ almost-sure inequality, valid for any $-\infty \leq m < m' \leq 1$ and uniform in θ:

$$|W_{m't}^e(\theta) - W_{mt}^e(\theta)| < K\rho^{t-m'} \tag{14}$$

The proof follows the same lines of step 1 in the score section (section 2.3.2). The only difference is that we cannot use $\mathbb{E}_{\theta^\star}\left[\cdot|Y_{m:t+1}\right] \xrightarrow{\theta^\star \ as} \mathbb{E}_{\theta^\star}\left[\cdot|Y_{-\infty:t+1}\right]$ when dealing with $E_\theta\left[\cdot|Y_{m:t+1}\right]$ instead of $\mathbb{E}_{\theta^\star}\left[\cdot|Y_{m:t+1}\right]$. We have to first use Cauchy-ness to take the limit and then extend the geometric bound of interest to the limit, exactly similar to what is done in step 1 in the log-likelihood section (section 2.3.1) and in the next step (step 1.2) for W_{mt}^v. The proof is omitted for brevity.

Step 1.2: $W_{mt}^v(\theta)$ is $\theta^\star$ almost surely uniform Cauchy

Similar to step 1 in the log-likelihood section (section 2.3.1), step 1 in the score section (section 2.3.2) and step 1.1 for $W_{mt}^e(\theta)$, we define $W_t^v(\theta) = W_{-\infty t}^v(\theta)$ and show the following $\theta^\star$ almost-sure inequality, valid for any $-\infty \leq m < m' \leq 1$ and uniform in θ:

$$|W_{m't}^v(\theta) - W_{mt}^v(\theta)| < K\rho^{t-m'} \tag{15}$$

The following $\theta^\star$ almost-sure bounds hold for any $-\infty < m \leq m' \leq r \leq s \leq t$, uniformly in θ:

$$|Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]| \leq K\rho^{r-m'} \tag{16}$$

$$|Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]| \leq K\rho^{t-s} \tag{17}$$

$$|Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t}\right]| \leq K\rho^{s-r} \tag{18}$$

The proofs of the bounds (16), (17) and (18) follow from the merging properties along the same lines of bounds (9), (10), (11) and (12) in the score section (section 2.3.2) and bound (4) in the log-likelihood section (section 2.3.1). They are omitted for brevity.

Note that for any $-\infty < m < m' \leq 1$:

$$\begin{aligned} W^v_{m't}(\theta) - W^v_{mt}(\theta) = &\left(\sum_{r=m'}^{t}\sum_{s=m'}^{t} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t+1}\right] - \sum_{r=m'}^{t-1}\sum_{s=m'}^{t-1} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t}\right]\right) \\ &- \left(\sum_{r=m}^{t}\sum_{s=m}^{t} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right] - \sum_{r=m}^{t-1}\sum_{s=m}^{t-1} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]\right) \end{aligned} \tag{19}$$

The bounds (16), (17) or (18) apply to different ways of grouping the terms in the sum (19). The idea is to split the sums into different regions, and then apply the sharpest bound available in each region. This is what was done explictly in the log-likelihood and score sections; the "region management" becomes too cumbersome here. We use a higher level approach. Fix m,m' and t and partition $A = m \leq r \leq t, m \leq s \leq t$ as $A_1 \cup A_2 \cup A_3 \cup A_4$ where:

$$\begin{aligned} A_1 &= \{m \leq r < m', m \leq s \leq t-1\} \cup \{m \leq r \leq t-1, m \leq s < m'\} \\ A_2 &= \{m' \leq r \leq t-1, m' \leq s \leq t-1\} \\ A_3 &= \{r = t, m' \leq s \leq t\} \cup \{m' \leq s \leq t, s = t\} \\ A_4 &= \{r = t, m \leq s < m'\} \cup \{m \leq r < m', s = t, \} \end{aligned}$$

Note that:

$$\begin{aligned}
W^v_{m't}(\theta) - W^v_{mt}(\theta) = &\sum_{A_1} \left(Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]\right) \\
&+ \sum_{A_2} \left(\left(Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t}\right]\right)\right. \\
&\qquad \left. - \left(Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]\right)\right) \\
&+ \sum_{A_3} \left(Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right]\right) \\
&+ \sum_{A_4} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right]
\end{aligned}$$

By the bounds (16), (17) and (18):

$$\begin{aligned}
|W^v_{m't}(\theta) - W^v_{mt}(\theta)| \leq &\sum_{A_1} K\rho^{t-r\vee s} \wedge \rho^{r\vee s - r\wedge s} + \sum_{A_2} K\rho^{t-r\vee s} \wedge \rho^{r\vee s - r\wedge s} \wedge \rho^{r\wedge s - m'} \\
&+ \sum_{A_3} K\rho^{r\vee s - r\wedge s} \wedge \rho^{r\wedge s - m'} + \sum_{A_4} K\rho^{r\vee s - r\wedge s}
\end{aligned}$$

Furthermore:

- On A_1, $r \wedge s - m' \leq 0 \leq r \vee s - r \wedge s$.
- On A_3, $t - r \vee s = 0 \leq r \vee s - r \wedge s$.
- On A_4, $t - r \vee s = 0 \leq r \vee s - r \wedge s$ and $r \wedge s - m' \leq 0 \leq r \vee s - r \wedge s$.

So that:

$$\begin{aligned}
|W^v_{m't}(\theta) - W^v_{mt}(\theta)| &\leq K \sum_{A_1 \cup A_2 \cup A_3 \cup A_4} \rho^{t-r\vee s} \wedge \rho^{r\vee s - r\wedge s} \wedge \rho^{r\wedge s - m'} \\
&= K \sum_{r=m}^{t} \sum_{s=m}^{t} \rho^{(t-r\vee s)\vee(r\vee s - r\wedge s)\vee(r\wedge s - m')}
\end{aligned}$$

Define $n = t - m + 1$, $\rho_n = \rho^n$, $a = \frac{m'-m+1}{t-m+1}$ and the function $g(r,s) := (1 - r \vee s) \vee$

$(r \vee s - r \wedge s) \vee (r \wedge s - a)$ for $(r,s) \in [0,1] \times [0,1]$. Then:

$$\begin{aligned}
&\sum_{r=m}^{t} \sum_{s=m}^{t} \rho^{(t-r\vee s)\vee(r\vee s - r\wedge s)\vee(r\wedge s - m')} \\
&\quad = n^2 \frac{1}{n^2} \sum_{r=1}^{n} \sum_{s=1}^{n} a^{g\left(\frac{r}{n},\frac{s}{n}\right)} \\
&\quad \leq n^2 \int_{0\leq r\leq 1} \int_{0\leq s\leq 1} a^{g(r,s)} dr ds \qquad \text{because } x \to a^x \text{ is decreasing} \\
&\quad = n^2 6 \frac{3a^{1/3} - 4a^{1/2} + a}{\log^2 a} \\
&\quad \leq n^2 6 \frac{3a^{1/3}}{\log^2 a} \qquad \text{because } a - 4\sqrt{a} < 0 \text{ when } 0 < a < 1 \\
&\quad = \frac{18K}{\rho} n^2 \frac{1}{n^2 \log^2 \rho} \rho^{n/3} \\
&\quad \leq \frac{18K}{\rho \log^2 \rho} \left(\rho^{1/3}\right)^{t-m'}
\end{aligned}$$

This proves (15) for $-\infty \leq m < m' \leq 1$. As a consequence $W^v_{mt}(\theta)$ is $\theta^\star$ almost surely a uniform Cauchy sequence, and as such converges to a limit $W^v_t(\theta) = W^v_{\infty t}(\theta)$ as $m \to -\infty$. (15) extends to $m = -\infty$ by continuity.

In order to give an explicit "infinite-past" representation for $W^v_t(\theta)$ which will be used to apply the ergodic theorem in step 3.2, note that (16) implies that:

$$Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{m:t+1}\right] \xrightarrow{\theta^\star \ as} Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{-\infty:t+1}\right]$$

Then bounds (16), (17) and (18) extend to $m = -\infty$ and imply that $Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{-\infty:t+1}\right] - Cov_\theta \left[J_r(\theta), J_s(\theta) | Y_{-\infty:t}\right]$ is (doubly) absolutely summable in $r \to -\infty$ and $s \to -\infty$,

and $Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{-\infty:t+1}\right]$ is absolutely summable in $s \to -\infty$. Rewrite W^v_{mt}:

$$\begin{aligned} W^v_{mt}(\theta) &= \sum_{r=m'}^{t}\sum_{s=m'}^{t} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t+1}\right] - \sum_{r=m'}^{t-1}\sum_{s=m'}^{t-1} Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m':t}\right] \\ &= \sum_{s=m}^{t} Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{m:t+1}\right] + \sum_{s=m}^{t-1} Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{m:t+1}\right] \\ &\quad + \sum_{r=m}^{t-1}\sum_{s=m}^{t-1}\left(Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{m:t}\right]\right) \end{aligned}$$

By taking m to $-\infty$ we get the "infinite-past" representation of $W^v_{mt}(\theta)$:

$$\begin{aligned} W^v_t(\theta) = W^v_{-\infty t}(\theta) = &\sum_{s=-\infty}^{t} Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{-\infty:t+1}\right] + \sum_{s=-\infty}^{t-1} Cov_\theta\left[J_s(\theta), J_t(\theta)|Y_{-\infty:t+1}\right] \\ &+ \sum_{r=-\infty}^{t-1}\sum_{s=-\infty}^{t-1}\left(Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{-\infty:t+1}\right] - Cov_\theta\left[J_r(\theta), J_s(\theta)|Y_{-\infty:t}\right]\right) \end{aligned}$$

Step 1.3: $W^c_{mt}(\theta)$ is $\theta^\star$ almost surely uniform Cauchy

Similar to step 1 in the log-likelihood section (section 2.3.1), step 1 in the score section (section 2.3.2), step 1.1 for $W^e_{mt}(\theta)$ and step 1.2 for $W^v_{mt}(\theta)$. We define $W^c_t(\theta) = W^c_{-\infty t}(\theta)$ and show the following $\theta^\star$ almost-sure inequality, valid for any $-\infty \le m < m' \le 1$ and uniform in θ:

$$|W^c_{m't}(\theta) - W^c_{mt}(\theta)| < K\rho^{t-m'} \tag{20}$$

Along the same lines as steps 1.1 and 1.3 and omitted for brevity.

Step 2.1: $\left|\frac{1}{T}\sum_{t=1}^{T-1} W^e_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^e_t(\theta)\right| \xrightarrow{\theta^\star \ as} 0.$

Step 2.2: $\left|\frac{1}{T}\sum_{t=1}^{T-1} W^v_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^v_t(\theta)\right| \xrightarrow{\theta^\star \ as} 0.$

Step 2.3: $\left|\frac{1}{T}\sum_{t=1}^{T-1} W^c_{1t}(\theta) - \frac{1}{T}\sum_{t=1}^{T-1} W^c_t(\theta)\right| \xrightarrow{\theta^\star \ as} 0.$

These three steps follow from the geometric bounds (14), (15) and (20) similarly to step 2 in the log-likelihood section (section 2.3.1) and step 2 in the score section (section 2.3.2).

Step 3.1: $\left\|\frac{1}{T}\sum_{t=1}^{T-1} W^e_{1t}(\theta) - i^e(\theta)\right\| \xrightarrow{\theta^\star \ as} 0.$

Step 3.2: $\left\|\frac{1}{T}\sum_{t=1}^{T-1} W^v_{1t}(\theta) - i^v(\theta)\right\| \xrightarrow{\theta^\star \ as} 0.$

Step 3.3: $\left\| \frac{1}{T} \sum_{t=1}^{T-1} W_{1t}^c(\theta) - i^c(\theta) \right\| \xrightarrow{\theta^\star \ as} 0$.

These two steps follow by the functional ergodic theorem. Their infinite-past representations show that $W_{1t}^e(\theta)$, $W_{1t}^v(\theta)$ and $W_{1t}^c(\theta)$ are ergodic, and continuity follows from the uniform limits. The functional ergodic theorem applies exactly as in step 3 of the log-likelihood section (section 2.3.1).

Note that $i(\theta^\star) = I$, for the usual reason, namely:

$$\begin{aligned}
Ti_T(\theta) &= -\nabla_\theta^2 \log P_\theta(Y_{2:T}|Y_1) \\
&= -\frac{\nabla_\theta^2 P_\theta(Y_{2:T}|Y_1)}{P_\theta^2(Y_{2:T}|Y_1)} + \frac{\nabla_\theta P_\theta(Y_{2:T}|Y_1) \nabla_\theta' P_\theta(Y_{2:T}|Y_1)}{P_\theta^2(Y_{2:T}|Y_1)} \\
T\mathbb{E}_{\theta^\star}[i_T(\theta^\star)|Y_1] &= -\underbrace{\sum_{y_{2:T}} \nabla_{\theta^\star}^2 P_{\theta^\star}(y_{2:T}|Y_1)}_{=\nabla^2 \sum = \nabla^2 1 = 0} + T^2 \mathbb{E}_{\theta^\star}\left[s_T s_T'|Y_1\right] \\
\mathbb{E}_{\theta^\star}[i_T(\theta^\star)] &\to i(\theta^\star) \\
\text{and } \mathbb{E}_{\theta^\star}[i_T(\theta^\star)] &= T\mathbb{E}_{\theta^\star}[s_T s_T'] \to I
\end{aligned}$$

2.4 Limit theorems under non-stationarity

Under stationarity, I proved a uniform law of large numbers for the log-likelihood, a central limit theorem for the score and a uniform law of large numbers for the observed information in 2.3. Local asymptotic normality and the asymptotic distribution of the maximum likelihood estimator will follow under standard arguments, but we would like those results to hold under nonstationarity too, when the true initial distribution $\mu^\star$ is different from the stationary one $\mu^\diamond$. In this section I show how to extend the central limit theorem to the nonstationary case. Likewise, the two uniform laws of large numbers hold under nonstationarity.

2.4.1 Preliminary lemma: the score has bounded differences

Let σ_T^i be the i^{th} coefficient of the unscaled score Ts_T. σ_T^i has bounded differences uniformly in T and t, meaning that there is c_i such that for any $y_{1:T}$, any $\hat{y}_t$:

$$\sigma_T^i\left(y_1, \ldots, y_{t-1}, y_t, y_{t+1}, \ldots, y_T\right) - \sigma_T^i\left(y_1, \ldots, y_{t-1}, \hat{y}_t, y_{t+1}, \ldots, y_T\right) \leq c_i$$

Proof. This follows from almost-sure Cauchy bounds proved above, specifically (8). Assume θ is scalar for ease of notation. Remember that we wrote $s_T(y_{1:T}) = \frac{1}{T}\sum_{t=1}^{T-1} V_{1t}(y_{1:t+1})$ and that (8) says that for any $m \leq m' \leq t$, $|V_{m't} - V_{mt}| \leq K\rho^{t-m'}$ almost-surely. Write $y_{1:t} = (y_1, \ldots, y_{s-1}, y_s, y_{s+1}, \ldots, y_t)$ and $\hat{y}_{1:t} = (y_1, \ldots, y_{s-1}, \hat{y}_s, y_{s+1}, \ldots, y_t)$. We have:

$$\begin{aligned}
|\sigma_T(y_{1:T}) - \sigma_T(\hat{y}_{1:T})| &= \left|\sum_{t=1}^{T-1} V_{1t}(y_{1:t+1}) - V_{1t}(\hat{y}_{1:t+1})\right| \\
&\leq \left|\sum_{t=1}^{T-1} V_{1t}(y_{1:t+1}) - V_{s+1t}(y_{s+1:t+1})\right| + \left|\sum_{t=1}^{T-1} V_{s+1t}(y_{s+1:t+1}) - V_{1t}(\hat{y}_{1:t+1})\right| \\
&\leq 2\frac{K}{1-\rho}
\end{aligned}$$

□

2.4.2 Nonstationary central limit theorem for the score

In this section I give a proof of the non-stationary central limit theorem for the score using merging and the bounded differences property of the score. The argument may be of more general interest. It is heuristically appealing: observations are distributed more and more according to the stationary distribution, and no single observation has an overwhelming influence on the score. The argument may be of more general interest. It can be used to show non-stationary central limit theorems for Markov chains or other merging processes such as hidden Markov models. One advantage is that it can handle non-additively-separable functions in addition to the more usual averages $\frac{1}{T}\sum_{t=1}^{T} f(Y_t)$.

Let $\tilde{Y}_t$ be the sequence under stationarity and Y_t under any initial distribution. Write $Y = Y_{1:T}$, $\tilde{Y} = \tilde{Y}_{1:T}$, $Y_{-t} = Y_{1:t-1,t+1,T}$ and $\tilde{Y}_{-t} = \tilde{Y}_{1:t-1,t+1,T}$. $\tilde{Y}$ and Y do not have to live on the same probability space; this will be made more precise below. Let $s_T = \frac{1}{T}\sigma_T(Y_{1:T})$ be the score for the non-stationary model and $\tilde{s}_T = \frac{1}{T}\sigma_T\left(\tilde{Y}_{1:T}\right)$ the score for the stationary model, both computed under the potentially misspecified assumption that the data are generated with some arbitrary initial stationary distribution μ. In section 2.3.2 I have shown that, for any μ, the following central limit theorem for the score holds under stationarity:

$$\sqrt{T}\tilde{s}_T \Rightarrow \mathcal{N}(0, I)$$

Now I want to show:

$$\sqrt{T}s_T \Rightarrow \mathcal{N}(0, I)$$

Let $d_{TV}(Z_1, Z_2)$ be the notation for the total variation distance between the distributions of any two random variables Z_1 and Z_2.

Recall that Y satisfies merging, i.e., there is $\rho < 1$, $c > 0$ such that:

$$d_{TV}\left(Y_t, \tilde{Y}_t\right) < c\rho^t$$

Assume that s is scalar for simplicity.

For P_1 and P_2 probability distributions on $\mathbb{R}$, write $P_1 \otimes P_2$ for the space of measures on $\mathbb{R}^2$ whose marginals are P_1 and P_2 (not to be confused with the product measure $P_1 \times P_2$). Let W be the Wasserstein metric between probability distributions on $\mathbb{R}$:

$$W(P_1, P_2) = \inf_{P \in P_1 \otimes P_2} \left(\int (z_1 - z_2)^2 P(dz) \right)^{1/2}$$

It is well-known that W metrizes weak convergence for probability measures with finite second moments.

For any two random variables Z_1 and Z_2, let $W(Z_1, Z_2)$ be the notation for $W(P_{Z_1}, P_{Z_2})$. Then:

$$W(Z_1, Z_2) = \inf_{P \in P_{Z_1} \otimes P_{Z_2}} \mathbb{E}_P \left[(Z_1 - Z_2)^2 \right]^{1/2}$$

In particular: for $a > 0$:

$$W(aZ_1, aZ_2) = \inf_{P \in P_{Z_1} \otimes P_{Z_2}} \mathbb{E}_P \left[(aZ_1 - aZ_2)^2 \right]^{1/2} = aW(Z_1, Z_2)$$

Consider the following inequality where on each line $Y_{1:t-1}, Y_t, \tilde{Y}_t, \tilde{Y}_{t+1:T}$ must have any joint distribution respecting the marginal distributions of Y and $\tilde{Y}$, but these

joint distributions do not have to be compatible between lines:

$$\begin{aligned}
W\left(s_T, \tilde{s}_T\right) &= W\left(\frac{1}{T}\sigma_T\left(Y_1, \ldots, Y_T\right), \frac{1}{T}\sigma_T\left(\tilde{Y}_1, \ldots, \tilde{Y}_T\right)\right) \\
&= \frac{1}{T}W\left(\sigma_T\left(Y_1, \ldots, Y_T\right), \sigma_T\left(\tilde{Y}_1, \ldots, \tilde{Y}_T\right)\right) \\
&\leq \frac{1}{T}\Bigg(W\left(\sigma_T\left(Y_1, \ldots, Y_T\right), \sigma_T\left(Y_1, \ldots, Y_{T-1}, \tilde{Y}_T\right)\right) \\
&\quad + W\left(\sigma_T\left(Y_1, \ldots, Y_{T-1}, \tilde{Y}_T\right), \sigma_T\left(Y_1, \ldots, Y_{T-2}, \tilde{Y}_{T-1}, \tilde{Y}_T\right)\right) \\
&\quad + \ldots \\
&\quad + W\left(\sigma_T\left(Y_1, Y_2, \tilde{Y}_3, \ldots, \tilde{Y}_T\right), \sigma_T\left(Y_1, \tilde{Y}_2, \ldots, \tilde{Y}_T\right)\right) \\
&\quad + W\left(\sigma_T\left(Y_1, \tilde{Y}_2, \ldots, \tilde{Y}_T\right), \sigma_T\left(\tilde{Y}_1, \ldots, \tilde{Y}_T\right)\right)\Bigg)
\end{aligned}$$

Let us bound each term separately. Fix t and define:

$$h(y_t) = \sigma_T\left(Y_1, \ldots, Y_{t-1}, y_t, \tilde{Y}_{t+1}, \ldots, \tilde{Y}_T\right)$$

Fix $P \in P_Y \otimes P_{\tilde{Y}}$:

$$\begin{aligned}
\mathbb{E}_P\left[\left(h\left(Y_t\right) - h\left(\tilde{Y}_t\right)\right)^2\right] &\leq \mathbb{E}_P\left[c^2 1\left[Y_t \neq \tilde{Y}_t\right]\right] \qquad \text{by bounded differences of the score} \\
&= c^2 P\left(Y_t \neq \tilde{Y}_t\right)
\end{aligned}$$

And:

$$\begin{aligned}
W\left(h\left(Y_t\right), h\left(\tilde{Y}_t\right)\right)^2 &= \inf_{P \in P_Y \otimes P_{\tilde{Y}}} \mathbb{E}_P\left[\left(h\left(Y_t\right) - h\left(\tilde{Y}_t\right)\right)^2\right] \\
&\leq \inf_{P \in P_Y \otimes P_{\tilde{Y}}} c^2 P\left(Y_t \neq \tilde{Y}_t\right)
\end{aligned}$$

Looking only at the marginal at time-horizon t, it is well-known that there is $P_t^\star \in P_{Y_t} \otimes P_{\tilde{Y}_t}$ such that:

$$\inf_{P \in P_{Y_t} \otimes P_{\tilde{Y}_t}} P\left(Y_t \neq \tilde{Y}_t\right) = \min_{P \in P_{Y_t} \otimes P_{\tilde{Y}_t}} P\left(Y_t \neq \tilde{Y}_t\right) = P_t^\star\left(Y_t \neq \tilde{Y}_t\right) = d_{TV}\left(Y_t, \tilde{Y}_t\right)$$

We want to extend this property to the joint probabilities on $1:T$. Fix $P_t^\star$ as above. Let $P^\star \in P_Y \otimes P_{\tilde{Y}}$ such that $\left(Y_t^\star, \tilde{Y}_t^\star\right) \sim P_t^\star$, $Y_{-t}^\star$ and $\tilde{Y}_{-t}^\star$ are independent

conditionally on $\left(Y_t^\star, \tilde{Y}_t^\star\right)$, $Y_{-t}^\star | Y_t^\star \sim P_{Y_{-t}|Y_t}$ and $\tilde{Y}_{-t}^\star | \tilde{Y}_t^\star \sim P_{\tilde{Y}_{-t}|\tilde{Y}_t}$. Then:

$$P^\star\left(Y_t \neq \tilde{Y}_t\right) = P_t^\star\left(Y_t \neq \tilde{Y}_t\right) = d_{TV}\left(Y_t, \tilde{Y}_t\right)$$

(Although we don't need it for bounding W, in fact $P^\star$ achieves $\inf_{P \in P_Y \otimes P_{\tilde{Y}}} P\left(Y_t \neq \tilde{Y}_t\right)$ because for any P, $d_{TV}\left(Y_t, \tilde{Y}_t\right) \leq P\left(Y_t \neq \tilde{Y}_t\right)$.)

As a consequence:

$$W\left(h\left(Y_t\right), h\left(\tilde{Y}_t\right)\right)^2 \leq c^2 \; d_{TV}\left(Y_t, \tilde{Y}_t\right)$$

And thanks to merging:

$$W\left(h\left(Y_t\right), h\left(\tilde{Y}_t\right)\right) \leq c\sqrt{\rho}^t$$

Putting all the terms back together:

$$W\left(s_T, \tilde{s}_T\right) \leq \frac{1}{T} c(1 + \sqrt{\rho} + \ldots + \sqrt{\rho}^T) \leq \frac{1}{T} c \frac{1}{1 - \sqrt{\rho}}$$

So that finally:

$$W\left(\sqrt{T} s_T, \mathcal{N}(0, I)\right) \leq W\left(\sqrt{T} s_T, \sqrt{T} \tilde{s}_T\right) + W\left(\sqrt{T} \tilde{s}_T, \mathcal{N}(0, I)\right) \to 0$$

$$\text{ie } \sqrt{T} s_T \Rightarrow \mathcal{N}(0, I)$$

2.5 Proof of Theorem 3 (uniform LAN) and Theorem 4 (asymptotic distribution of the MLE)

(Uniform) local asymptotic normality (Theorem 3) is a standard consequence of the central limit theorem for the score and the uniform law of large numbers for the observed information (see van der Vaart (1998)).

We show that $\theta^\star$ is the unique maximum of $L(\theta) = \mathbb{E}_{\theta^\star}[\log P_\theta\left(Y_1 | Y_{-\infty:0}\right)]$ in two steps.

Step 1: $P_\theta\left(Y_{1:T} | Y_{-\infty:m}\right) \xrightarrow{\theta^\star \; as} P_\theta(Y_{1:T})$ **when** $m \to -\infty$

We use the merging property of $z_t = (x_t, y_t)$ directly:

$$
\begin{aligned}
&|P_\theta(y_{1:T}) - P_\theta\left(y_{1:T}|y_{-\infty:m}\right)| \\
&\quad \leq \left| \sum_{x_0,y_0} P_\theta(y_{1:T}|x_0,y_0) P_\theta(x_0,y_0) - \sum_{x_0,y_0} P_\theta(y_{1:T}|x_0,y_0) P_\theta(x_0,y_0|y_{-\infty:m}) \right| \\
&\quad \leq \sum_{x_0,y_0} P_\theta(y_{1:T}|x_0,y_0) \left| \sum_{x_{m+1},y_{m+1}} P_\theta(x_0,y_0|x_{m+1},y_{m+1}) (P_\theta(x_{m+1},y_{m+1}) - P_\theta(x_{m+1},y_{m+1}|y_{-\infty:m})) \right| \\
&\quad \leq \sum_{x_0,y_0} P_\theta(y_{1:T}|x_0,y_0) c\rho^m \qquad \text{by merging} \\
&\quad \leq d_x d_y c \rho^m \\
&\quad \xrightarrow[m\to-\infty]{} 0
\end{aligned}
$$

Step 2: $\theta^\star$ is the unique maximum of $L(\theta)$

Let us show by contradiction that, for $\theta \neq \theta^\star$, $P_\theta\left(Y_1|Y_{-\infty:0}\right)$ is not $\theta^\star$ almost surely equal to $P_{\theta^\star}\left(Y_1|Y_{-\infty:0}\right)$. Suppose it is. Then by the law of iterated expectations and stationarity, $P_\theta\left(Y_{1:T}|Y_{-\infty:0}\right) = P_{\theta^\star}\left(Y_{1:T}|Y_{-\infty:0}\right)$ ($\theta^\star$-as) for any $T \geq 1$; and by integration and stationarity: $P_\theta\left(Y_{1:T}|Y_{-\infty:m}\right) = P_{\theta^\star}\left(Y_{1:T}|Y_{-\infty:m}\right)$ ($\theta^\star$-as) for any $T \geq 1 \geq m$. Then by step 1, $P_\theta\left(Y_{1:T}\right) = P_{\theta^\star}\left(Y_{1:T}\right)$ for any $T \geq 1$, which contradicts the identification assumption (A3).

Then by the strict Jensen inequality:

$$
\begin{aligned}
\mathbb{E}_{\theta^\star}\left[\log \frac{P_\theta\left(Y_1|Y_{-\infty:0}\right)}{P_{\theta^\star}\left(Y_1|Y_{-\infty:0}\right)}\right] &< \log \mathbb{E}_{\theta^\star}\left[\frac{P_\theta\left(Y_1|Y_{-\infty:0}\right)}{P_{\theta^\star}\left(Y_1|Y_{-\infty:0}\right)}\right] \\
&= \log \mathbb{E}_{\theta^\star}\left[\mathbb{E}_{\theta^\star}\left[\frac{P_\theta\left(Y_1|Y_{-\infty:0}\right)}{P_{\theta^\star}\left(Y_1|Y_{-\infty:0}\right)}\middle| Y_{-\infty:0}\right]\right] \\
&= 0
\end{aligned}
$$

Thus:

$$L(\theta) < L\left(\theta^\star\right)$$

The continuity of L and the compactness of Θ imply that $\theta^\star$ is a well-separated maximum and the uniform law of large numbers for L_T implies strong consistency. I is invertible by identification and our smoothness assumptions. Asymptotic normality

of the maximum likelihood estimator is a standard consequence of the uniform local asymptotic normality property and consistency.

2.6 Asymptotic distribution of the Bayesian posterior: Bernstein–von Mises theorem (proof of Theorem 5)

I apply the weakly dependent Bernstein–von Mises of Connault (2014).

In a hidden Rust model, the domination assumption (Assumption 1 in that paper) is verified. Local asymptotic normality (Assumption 4) is of course our Theorem 3. The prior support assumption (Assumption 2) is (A5) here. Assumptions 3 (uniformly consistent tests), 5 (a local linear lower bound for the score), 6 (a large deviation inequality for the score) and 7 (a large deviation inequality for blocks of data) remain to be checked.

2.6.1 Uniformly consistent estimators

Uniformly consistent estimators can be used to build uniformly consistent tests. By uniformly consistent estimators, I mean estimators $\hat{\theta}_T$ such that, for some distance d:

$$\forall \epsilon, \qquad \sup_{\theta} P_\theta \left(d\left(\hat{\theta}_T, \theta\right) \geq \epsilon \right) \to 0$$

d can be any distance as long as it is locally stronger than the reference Euclidean distance around $\theta^\star$, i.e.:

$$\forall \epsilon > 0, \exists \eta > 0 : \qquad \|\theta - \theta^\star\| > \epsilon \Longrightarrow d\left(\theta, \theta^\star\right) > \eta$$

See section 2.6.2 for why we need d to be locally stronger than the Euclidean distance.

$\hat{\theta}_T$ is not an estimator that is meant to be used in practice. It is used only as a technical device in the proof of the Bernstein–von Mises theorem. It does not matter if $\hat{\theta}_T$ has terrible short-sample properties or is not statistically efficient, as long as it is uniformly consistent. I will appeal to the generic identification structure theorem of this paper (Theorem 2) to exhibit such a $\hat{\theta}_T$.

By assumption (A3), the model is identified under stationarity. By merging, this implies it is also identified under a different initial distribution (any marginal can be arbitrarily well approximated by waiting long enough). By Theorem 2, let T_0 be a time horizon such that the marginal stationary distribution of T_0 consecutive y's identify θ. Let π be the corresponding marginal, i.e., the distribution of $y_{1:T_0}$ under stationarity. Let $(\hat{x}_s, \hat{y}_s)$ be non-overlapping blocks of T_0 consecutive (x_t, y_t)'s and $S = \lfloor T/T_0 \rfloor$. Let $\hat{\theta}_T = \hat{\pi}_T$ be the empirical distribution estimator of π defined by:

$$\hat{\pi}_T \left(Y_{1:T}; \hat{y} = y_{1:T_0} \right) = \frac{1}{S} \sum_{s=1}^{S} 1 \left[\hat{Y}_s = \hat{y} \right]$$

Finally let d_{TV} be the total variation distance. Because the model is identified, $d_{TV}\left(\hat{\pi}_T, \pi\right)$ is a particular choice of distance $d\left(\hat{\theta}_T, \theta\right)$. I show that $\hat{\pi}_T$ is a uniformly consistent estimator:

$$\forall \epsilon, \qquad \sup_{\theta} P_\theta \left(d_{TV}\left(\hat{\pi}_T, \pi\right) \geq \epsilon \right) \to 0 \tag{21}$$

To prove (21) we show a Dvoretzky-Kiefer-Wolfowitz type inequality, that is a quantitative bound going to zero with the time-series length: there is a sequence $\alpha(T) \xrightarrow[T \to \infty]{} 0$ such that:

$$\forall \theta, \qquad P_\theta \left(d_{TV}\left(\hat{\pi}_T, \pi\right) \geq \epsilon \right) \leq \alpha(T) \tag{22}$$

(22) follows in turn from a concentration bound for $d_{TV}\left(\hat{\pi}_T, \pi\right)$ around its expectation, and a separate bound for its expectation.

Step 1: concentration bound for $d_{TV}\left(\hat{\pi}_T, \pi\right)$.

We want to apply the concentration inequality (2). Let us check that $d_{TV}\left(\hat{\pi}_T, \pi\right)$ verifies a suitable bounded differences condition. By definition:

$$d_{TV}\left(\hat{\pi}_T, \pi\right) = \frac{1}{2} \sum_{\hat{y}} \left| \hat{\pi}_T\left(\hat{y}\right) - \pi\left(\hat{y}\right) \right|$$

For any sequence of observations $y_{1:T}$ and $\tilde{y}_{1:T}$:

$$\begin{aligned}
&d_{TV}\left(\hat{\pi}_T\left(y_{1:T}\right),\pi\right) - d_{TV}\left(\hat{\pi}_T\left(\tilde{y}_{1:T}\right),\pi\right) \\
&\quad\leq \frac{1}{2}\sum_{\hat{y}}\left|\hat{\pi}_T\left(y_{1:T};\hat{y}\right) - \hat{\pi}_T\left(\tilde{y}_{1:T};\hat{y}\right)\right| \qquad \text{by triangle inequality} \\
&\quad\leq \frac{1}{2}\frac{1}{S}\sum_{s=1}^{S} 1\left[\hat{y}_s \neq \hat{\tilde{y}}_s\right] \\
&\quad\leq \frac{1}{2}\frac{1}{S}\sum_{t=1}^{T_0} 1\left[y_t \neq \tilde{y}_t\right] \\
&\quad\leq \frac{1}{2}\frac{1}{T-T_0}\sum_{t=1}^{T} 1\left[y_t \neq \tilde{y}_t\right]
\end{aligned}$$

Thus $d_{TV}\left(\hat{\pi}_T,\pi\right)$ verifies a bounded differences condition with $c_t = \frac{1}{2}\frac{1}{T-T_0}$ and we can apply (2) as announced:

$$\begin{aligned}
P_{\theta,\mu}\left(d_{TV}\left(\hat{\pi}_T,\pi\right) > \mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right] + u\right) &\leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_x(\theta)\sum_{t=1}^{T}\left(\frac{1}{2}\frac{1}{T-T_0}\right)^2}\right) \\
&\leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_x\frac{T}{4(T-T_0)^2}}\right)
\end{aligned}$$

Step 2: bounding $\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]$.

To bound the expectation under an arbitrary non-stationary initial distribution I show a bound under the stationary distribution and I take it to the non-stationary case using merging properties. This is similar to the way I show a non-stationary central limit theorem for the score from a stationary central limit theorem using merging, in the local asymptotic normality section (section 2.5).

Step 2.1: bounding $\mathbb{E}_{\theta}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right] := \mathbb{E}_{\theta,\mu^\diamond(\theta)}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]$.

Paulin (2014) gives a bound under stationarity for the empirical distribution estimator of the one-dimensional marginal of a Markov chain (bound (3.31) p.21). We can apply this bound to the block Markov chain $\left(\hat{x},\hat{y}\right)_s$. Write λ for the joint distribution of $\left(\hat{x},\hat{y}\right)_1 = (x,y)_{1:T_0}$ under stationarity and $\hat{\lambda}_T$ for the corresponding empirical

distribution estimator. A consequence of (3.31) from Paulin (2014) is that there is a constant K such that, for T big enough:

$$\mathbb{E}_\theta\left[d_{TV}\left(\hat{\lambda}_T,\lambda\right)\right]\leq\sqrt{\frac{1}{S}}\sum_{\hat{y}}\sqrt{\lambda\left(\hat{y}\right)}$$

((3.31) in Paulin (2014) involves a technical quantity, the inverse of the "pseudo spectral gap" of the Markov chain $(\hat{x},\hat{y})$ $\gamma_{(\hat{x},\hat{y})}(\theta)$, which is shown to be bounded by the mixing time which in turn I showed to have a uniform bound in section 2.2.2.) Let $\hat{d}=d_{\hat{y}}=d_y^{T_0}$. Remember the general inequality between ℓ^p norms:

$$\left\|\lambda\right\|_{1/2}\leq\hat{d}^{\frac{1}{1/2}-\frac{1}{1}}\left\|\lambda\right\|_1=\hat{d}$$

So that:

$$\mathbb{E}_\theta\left[d_{TV}\left(\hat{\lambda}_T,\lambda\right)\right]\leq\sqrt{\frac{\hat{d}K}{T-T_0}}$$

Now consider the projection function $h_y\left(\hat{x},\hat{y}\right)=\hat{y}$, which takes the joint Markov chain $\left(\hat{x},\hat{y}\right)_s$ to its observable component $\hat{y}_s$. Then $\hat{\pi}_T$ and π are the distributions of $h_y\left(\hat{x},\hat{y}\right)$ under $\hat{\lambda}_T$ and λ, respectively (i.e., $\hat{\pi}_T=\hat{\lambda}_T\circ h_y^{-1}$ and $\hat{\pi}_T=\hat{\lambda}_T\circ h_y^{-1}$). Now if $d_{TV}(Z_1,Z_2)$ means the total variation distance between the distributions of two arbitrary random variables Z_1 and Z_2 and h is any function, d_{TV} satisfies $d_{TV}(h(Z_1),h(Z_2))\leq d_{TV}(Z_1,Z_2)$. In the case at hand:

$$\mathbb{E}_\theta\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]\leq\mathbb{E}_\theta\left[d_{TV}\left(\hat{\lambda}_T,\lambda\right)\right]\leq\sqrt{\frac{\hat{d}K}{T-T_0}}$$

Step 2.2: bounding $\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]$.

Remember that Y_t satisfies merging uniformly in θ and μ by assumption: there is $\rho<1$, $c>0$ such that:

$$d_{TV}(Y_t,\mu^\diamond)\leq c\rho^t$$

I show that $\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]-\mathbb{E}_\theta\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]$ goes to zero using the bounded differences property of $d_{TV}\left(\hat{\pi}_T,\pi\right)$ together with merging. Let $(Y_t)_t$ be distributed according to θ and μ (nonstationary) and $\left(\tilde{Y}_t\right)_t$ be distributed according to θ and stationary. Y and $\tilde{Y}$ do not have to live on the same probability space. Consider the following

inequality where on each line $Y_{1:t-1}, Y_t, \tilde{Y}_t, \tilde{Y}_{t+1:T}$ must have any joint distribution respecting the marginal distributions of Y and $\tilde{Y}$, but these joint distributions do not have to be compatible between lines:

$$
\begin{aligned}
&\left|\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right] - \mathbb{E}_{\theta}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right]\right| \\
&\quad = \left|\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_1, \ldots, Y_T\right), \pi\right)\right] - \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(\tilde{Y}_1, \ldots, \tilde{Y}_T\right), \pi\right)\right]\right| \\
&\quad \leq \left|\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_1, \ldots, Y_T\right), \pi\right)\right] - \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_1, \ldots, Y_{T-1}, \tilde{Y}_T\right), \pi\right)\right]\right| \\
&\qquad + \left|\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_1, \ldots, Y_{T-1}, \tilde{Y}_T\right), \pi\right)\right] - \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_1, \ldots, Y_{T-2}, \tilde{Y}_{T-1}, \tilde{Y}_T\right), \pi\right)\right]\right| \\
&\qquad + \ldots \\
&\qquad + \left|\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_1, Y_2, \tilde{Y}_3, \ldots, \tilde{Y}_T\right), \pi\right)\right] - \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_1, \tilde{Y}_2, \ldots, \tilde{Y}_T\right), \pi\right)\right]\right| \\
&\qquad + \left|\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_1, \tilde{Y}_2, \ldots, \tilde{Y}_T\right), \pi\right)\right] - \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(\tilde{Y}_1, \ldots, \tilde{Y}_T\right), \pi\right)\right]\right|
\end{aligned}
$$

Let us bound each term separately. Fix t, $1 \leq t \leq T$.

$$
\begin{aligned}
&\left|\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_{1:t-1}, Y_t, \tilde{Y}_{t+1:T}\right), \pi\right)\right] - \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_{1:t-1}, \tilde{Y}_t, \tilde{Y}_{t+1:T}\right), \pi\right)\right]\right| \\
&\quad \leq \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_{1:t-1}, Y_t, \tilde{Y}_{t+1:T}\right), \hat{\pi}_T\left(Y_{1:t-1}, \tilde{Y}_t, \tilde{Y}_{t+1:T}\right)\right)\right] \qquad \text{by triangle inequality} \\
&\quad = \mathbb{E}\left[\frac{1}{S} 21\left[Y_t \neq \tilde{Y}_t\right]\right]
\end{aligned}
$$

Exactly as in section 2.5, we can build $P^\star$ such that $\left(Y, \tilde{Y}\right) \sim P^\star$ and $P^\star\left(Y_t \neq \tilde{Y}_t\right) = d_{TV}\left(Y_t, \tilde{Y}_t\right)$. Hence the bound:

$$
\begin{aligned}
&\left|\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_{1:t-1}, Y_t, \tilde{Y}_{t+1:T}\right), \pi\right)\right] - \mathbb{E}\left[d_{TV}\left(\hat{\pi}_T\left(Y_{1:t-1}, \tilde{Y}_t, \tilde{Y}_{t+1:T}\right), \pi\right)\right]\right| \\
&\quad \leq \frac{2}{T - T_0} d_{TV}\left(Y_t, \tilde{Y}_t\right) \\
&\quad \leq \frac{2c}{T - T_0} \rho^t
\end{aligned}
$$

Putting back all the terms together, we get the bound:

$$
\left|\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right] - \mathbb{E}_{\theta}\left[d_{TV}\left(\hat{\pi}_T, \pi\right)\right]\right| \leq \frac{2c}{T - T_0}(1 + \rho + \ldots + \rho^T) \leq \frac{2c}{(T - T_0)(1 - \rho)}
$$

And with step 2.1:

$$\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right] \leq \frac{2c}{(T-T_0)(1-\rho)}\sqrt{\frac{\hat{d}K}{T-T_0}}$$

Finally, putting steps 1 and 2 together:

$$\begin{aligned} P_{\theta,\mu}\left(d_{TV}\left(\hat{\pi}_T,\pi\right)\geq\epsilon\right) &= P_{\theta,\mu}\left(d_{TV}\left(\hat{\pi}_T,\pi\right)\geq\mathbb{E}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]+\left(\epsilon-\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]\right)\right) \\ &\leq \exp\left(-\frac{1}{2}\frac{\left(\epsilon-\mathbb{E}_{\theta,\mu}\left[d_{TV}\left(\hat{\pi}_T,\pi\right)\right]\right)^2}{\tau_x\frac{T}{4(T-T_0)^2}}\right) \end{aligned}$$

This is of the form (22) (for a fixed ϵ and T big enough) and shows that $\hat{\theta}_T = \hat{\pi}_T$ is a uniformly consistent estimator as in (21).

2.6.2 Uniformly consistent tests (checking assumption (A3))

Let $\hat{\theta}_T$ be any uniformly consistent estimator in the sense of section 2.6.1 ($\hat{\pi}_T$ is such an estimator by section 2.6.1). Let $\epsilon > 0$. Recall that by definition of a uniformly consistent estimator (section 2.6.1), there is $\eta > 0$ such that:

$$\|\theta-\theta^\star\| > \epsilon \Longrightarrow d\left(\theta,\theta^\star\right) > \eta$$

Let us show that $\phi_T = 1\left[d\left(\hat{\theta}_T,\theta^\star\right)\geq\eta/2\right]$ is a uniformly consistent test for ϵ.

First,

$$\mathbb{E}_{\theta^\star}[\phi_T] = P_{\theta^\star}\left(d\left(\hat{\theta}_T,\theta^\star\right)\geq\eta/2\right)\to 0 \qquad \text{by consistency}$$

Second,

$$\begin{aligned} \mathbb{E}_\theta[1-\phi_T] &= P_\theta(\phi_T=0) \\ &= P_\theta\left(d\left(\hat{\theta}_T,\theta^\star\right)<\eta/2\right) \\ &\leq P_\theta\left(d\left(\theta,\theta^\star\right)-d\left(\hat{\theta}_T,\theta\right)<\eta/2\right) \qquad \text{by triangle inequality} \\ &= P_\theta\left(d\left(\hat{\theta}_T,\theta\right)>d\left(\theta,\theta^\star\right)-\eta/2\right) \end{aligned}$$

So that:

$$\begin{aligned}\sup_{\|\theta-\theta^\star\|>\epsilon} \mathbb{E}_\theta[1-\phi_T] &\leq \sup_{\|\theta-\theta^\star\|>\epsilon} P_\theta\left(d\left(\hat{\theta}_T,\theta\right)>\eta/2\right)\\ &\to 0 \qquad\qquad \text{by uniform consistency}\end{aligned}$$

2.6.3 Local linear lower bound for the score (checking assumption (A5))

We can rely on the smoothness of $\theta \to \mathbb{E}_\theta[s_T]$ on Θ. Note that as usual:

$$\begin{aligned}\nabla_{\theta^\star}\mathbb{E}_\theta[s_T] &= \sum_{\tilde{y}_{1:T}} s_T \nabla_{\theta^\star} P_\theta(\tilde{y}_{2:T}|\tilde{y}_1)P(\tilde{y}_1)\\ &= \sum_{\tilde{y}_{1:T}} s_T \frac{\nabla_{\theta^\star} P_\theta(\tilde{y}_{2:T}|\tilde{y}_1)}{P_{\theta^\star}(\tilde{y}_{2:T}|\tilde{y}_1)} P_{\theta^\star}(\tilde{y}_{2:T}|\tilde{y}_1)P(\tilde{y}_1)\\ &= T\mathbb{E}_{\theta^\star}[s_T s_T'] \\ &= \mathbb{E}_{\theta^\star}[i_T(\theta^\star)] \qquad \text{(see the end of section 2.3.3)}\end{aligned}$$

Write $h(\theta) = \nabla^2_\theta \mathbb{E}_\theta[s_T]$. Consider a second-order Taylor expansion around $\theta^\star$ with Lagrange remainder: there is $\bar{\theta}$, $\theta_i^\star \leq \bar{\theta}_i \leq \theta_i$, such that:

$$\mathbb{E}_\theta[s_T] = \mathbb{E}_{\theta^\star}[s_T] + \mathbb{E}_{\theta^\star}[i_T(\theta^\star)](\theta-\theta^\star) + (\theta-\theta^\star)' h\left(\bar{\theta}\right)(\theta-\theta^\star)$$

Since $\mathbb{E}_{\theta^\star}[i_T(\theta^\star)] \xrightarrow[T\to\infty]{} I$ (see section 2.3), I is invertible by assumption and h is bounded over Θ by smoothness and compactness, there is T_0, $\delta < 1$ and c such that for any $\|\theta-\theta^\star\| \leq \delta$, $T > T_0$:

$$\|\mathbb{E}_\theta[s_T] - \mathbb{E}_{\theta^\star}[s_T]\| \geq c\|\theta-\theta^\star\|$$

2.6.4 Large deviation inequality for the score (checking assumption (A6))

Let σ_T^i be the i^{th} coefficient of the unscaled score Ts_T. The score has bounded differences; see section 2.5. We can apply the concentration inequality (3): for any θ:

$$\begin{aligned}P_\theta\left(\left|\sigma_T^i/T - \mathbb{E}_\theta\left[\sigma_T^i/T\right]\right| > u\right) &\leq 2\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x(\theta)\sum_{t=1}^T (c_i/T)^2}\right)\\ &\leq 2\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x c_i/T}\right)\end{aligned}$$

To conclude, note that for a general random vector X:

$$P(\|X\|_2 > u) < d_X \max_i P\left(|X_i| > \frac{u}{\sqrt{d_X}}\right)$$

So that if $\bar{c} = \max_{1 \leq i \leq d_\theta} c_i$:

$$P_\theta\left(\|s_T - \mathbb{E}_\theta[s_T]\| > u\right) \leq 2\exp\left(-\frac{1}{2}\frac{u^2}{\tau_x \bar{c}/T}\right)$$

Thus assumption (A6) holds with $c = \tau_x \bar{c}$

2.6.5 Large deviation inequality for blocks (checking assumption (A7))

Let $R \in \mathbb{N}$ and define blocks $(\hat{x}_s, \hat{y}_s)$ to be non-overlapping blocks of R consecutive (x_t, y_t)'s. $(\hat{x}_s, \hat{y}_s)$ itself is a hidden Markov model and satisfies the concentration inequalities of Paulin (2014). In particular, for any g, $0 \leq g \leq 1$, applying the one-sided inequality (1) to $f(\hat{y}_1, \ldots, \hat{y}_S) = \frac{1}{S}\sum_{s=1}^S g_s$ where $g_s = g(\hat{y}_s)$ and we have:

$$\begin{aligned} P_\theta\left(\frac{1}{S}\sum_{s=1}^S g_s < \mathbb{E}\left[\frac{1}{S}\sum_{s=1}^S g_s\right] - u\right) &\leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_{\hat{x}}(\theta)/S}\right) \\ &\leq \exp\left(-\frac{1}{2}\frac{u^2}{\tau_{\hat{x}}/S}\right) \end{aligned}$$

Thus, assumption (A7) holds with $c_R = \tau_{\hat{x}}$.

References

Baum, L. E. and T. Petrie (1966): "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, 37, 1554–1563.

Connault, B. (2014): "A Weakly Dependent Bernstein–von Mises Theorem," *Working Paper*, http://economics.sas.upenn.edu/~connault/.

Decker, W., G.-M. Greuel, G. Pfister, and H. Schönemann (2015): "Singular 4-0-2 — A computer algebra system for polynomial computations," http://www.singular.uni-kl.de.

Douc, R., E. Moulines, and T. Ryden (2004): "Asymptotic Properties of the Maximum Likelihood Estimator in Autoregressive Models with Markov Regime," *The Annals of Statistics*, 32, 2254–2304.

Douc, R., E. Moulines, and D. Stoffer (2014): *Nonlinear time series: Theory, methods and applications*, CRC Press.

Fill, J. A. (1991): "Eigenvalue Bounds on Convergence to Stationarity for Nonreversible Markov Chains, with an Application to the Exclusion Process," *The Annals of Applied Probability*, 1, 62–87.

Krengel, U. (1985): *Ergodic Theorems*, Cambridge University Press.

Louis, T. (1982): "Finding the Observed Information Matrix when Using the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 44, 226–233.

Paulin, D. (2014): "Concentration Inequalities for Markov Chains by Marton Couplings and Spectral Methods," *Working Paper*, http://arxiv.org/abs/1212.2015v3.

Seneta, E. (2006): *Non-Negative Matrices and Markov Chains*, Springer.

Sommese, A. and C. Wampler (2005): *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*, World Scientific.

van der Vaart, A. (1998): *Asymptotic Statistics*, Cambridge University Press.