Asymptotic Stability of Nonlinear Filters in Hidden Markov Models and Its Applications

Tze Leung Lai

Stanford University

Outline

- Brief survey of previous works on asymptotic stability of the Kalman filter and nonlinear filters in HMMs
- HMM filter as a measure-valued Markov chain
- From V-uniform ergodicity (strong stability) of the underlying chain of HMM to that of HMM filter (as measure-valued chain)
- Modified Foster–Liapounov functions, coupling and regeneration
- Applications to parameter estimation in HMM, adaptive control, and sequential Monte Carlo

Filtering in Hidden Markov Models

- Unobserved Markov chain $\{x_n\}$ on state space $(\mathcal{X}, \mathcal{A})$ with transition probability density $p(x, x')$ w.r.t. ν.
- Observations y_n, taking values in $\mathcal{Y}$, are conditionally independent s.t. y_t has density function $g(\cdot|x_t)$ w.r.t. ν_Y.
- Filtering problem: Find posterior distribution $\pi_t = \mathcal{L}(x_t|y_1, \ldots, y_t)$.
- Optimal filter w.r.t. squared error loss: $E(x_t|y_1, \ldots, y_t)$.

Special Case: Kalman Filter

- $x_{n+1} = Fx_n + \gamma + w_n,\ y_n = Hx_n + \mu + \epsilon_n,$
  normal w_n, ϵ_n; normal initial distribution π_0 for x_0.

- π_t is normal $N(\widehat{x}_t, V_t)$, $\quad \widehat{x}_t = E_{\pi_0}(x_t | y_1, \ldots, y_t)$.

- $\widehat{x}_t = (F - K_t H)\widehat{x}_{t-1} + K_t(y_t - \mu) + \gamma$.
  K_t: Kalman gain matrix defined recursively.

Special Case: Kalman Filter (cont.)

- Suppose F is *asymptotically stable*, i.e., $\max_j |\lambda_j(F)| < 1$, where the $\lambda_j(F)$ are eigenvalues of F.
  - Then so is $F - KH$, where $K := \lim_{t\to\infty} K_t$ exists.
  - By asymptotic stability and the difference equation for $\widehat{x}_t$, V_t converges to its limit V exponentially fast.
  - Moreover, $\widehat{x}_t \xrightarrow{\mathcal{D}} \widehat{x}_\infty := E(x_0|y_0, y_{-1}, \ldots)$ exponentially fast, where E refers to expectation under the stationary distribution of $\{x_t\}$.

Special Case: Kalman Filter (cont.)

- Extension to general HMMs? *Asymptotic stability* in the sense that π_t forgets the "initial condition" π_0 exponentially fast?

- Key tools in Kalman filter case: π_t is normal. Hence need only show $V_t \to V$ and $\widehat{x}_t \xrightarrow{\mathcal{D}} \widehat{x}_\infty$ exponentially fast.
  Linear Kalman filter and quadratic Riccati recursions for $\widehat{x}_t$, V_t.

- General HMM: Dynamics of the Markov chain (x_t, π_t).
  - Analytic tools via measure-valued processes:

    Da Prato, Furhman and Malliavin (1995),

    Ocone and Pardoux (1996).

HMM Filter as Measure-Valued Markov Chain

$\mathcal{M}$: Space of probability measures on $\mathcal{X}$.
(x_t, π_t): Markov chain on $\mathcal{X} \times \mathcal{M}$.

- Transition from (x_{t-1}, π_{t-1}) to (x_t, π_t):
  - Use $p(x_{t-1}, \cdot)$ to generate x_t.
  - Use $g(\cdot|x_t)$ to generate y_t.
  - Define measure $\pi_t (\ll \nu)$ by

$$(*) \qquad \frac{d\pi_t}{d\nu}(x) = \frac{\int g(y_t|x)p(z,x)d\pi_{t-1}(z)}{\iint g(y_t|x')p(z,x')d\pi_{t-1}(z)d\nu(x')} \ .$$

- Transitions induced by $(*)$ are highly nonlinear, but numerator of $(*)$ corresponds to linear evolution: Three techniques to exploit the linearity in the literature ...

(A) Use **projective metrics** (e.g., Hilbert metric) to handle normalization in $(*)$, working with the linear evolution of its numerator.

Atar and Zeitouni (1997), Le Gland and Oudjane (2004).

- Hilbert metric on $\mathcal{M}^+(\mathcal{X})$ is defined by

$$h(\mu, \mu') = \log \frac{\sup_{A:\mu'(A)>0} \mu(A)/\mu'(A)}{\inf_{A:\mu'(A)>0} \mu(A)/\mu'(A)}$$

if μ and μ' are comparable, i.e., $\exists 0 < a \leq b$ s.t. $a\mu'(A) \leq \mu(A) \leq b\mu'(A)$ for all Borel sets A, with $h(\mu, \mu') = \infty$ otherwise. It is a projective metric, i.e., invariant under positive scalar multiplication.

(B) Use **bounds** on L_1-norms of $d\pi_t/d\nu - d\pi_t^0/d\nu$ in terms of certain **norms** of the difference of numerators of $(*)$.

Budhiraja and Ocone (1999).

(C) Use the **ergodic coefficient**

$$\alpha(K) = 1 - \sup_{x,x' \in \mathcal{X}, A \in \mathcal{A}} |K(x, A) - K(x', A)|$$

of n-step transition kernel K to bound the total variation norm of the signed measure $\pi_n(\mu) - \pi_n(\mu')$ via a Feyman–Kac semigroup of operators, where $\pi_n(\mu) =$ filter initialized at probability measure μ.

Del Moral and Guionnet (2001).

- The bounds in (B) assume the observation noise to be sufficiently small, so the HMM has "slightly hidden" states.

- Among other conditions, (A) and (C) require:

$$0 < \lambda_* \leq p(x, x') \leq \lambda^* < \infty \; \forall \; x, x' \in \mathcal{X} \quad \text{(uniformly recurrent } x_t\text{)}.$$

- *Chigansky and Lipster (2004)*: Another approach to weaken uniform recurrence to

$$\int_{\mathcal{X}} \{\operatorname{ess} \inf_{x' \in \mathcal{X}} p(x, x')\} d\pi^*(x) > 0, \; \sup_{x, x' \in \mathcal{X}} p(x, x') = \lambda^* < \infty,$$

where π^* is the stationary distribution of $\{x_t\}$ that is assumed to exist. But assumption still too strong for general state space, e.g., $\operatorname{ess\,inf}_{x' \in \mathcal{X}} p(x, x') = 0 \; \forall \; x$ when $p(x, x') = f(x - x') > 0$ with $f(\pm\infty) = 0$, as in Gaussian chains.

- A new approach: Replace uniform recurrence by V-uniform ergodicity of $\{x_t\}$ and use modified Liapounov–Foster functions.

  - Stochastic Liapounov function:

    $$E[V(x_t)|\mathcal{F}_{t-1}](= E_{x_{t-1}}[V(x_1)]) \leq V(x_{t-1}).$$

  - Foster–Liapounov function:

    $E_x V(x_1) \leq V(x) - \epsilon f(x) \quad \forall\, x \notin K$, for some measurable $f \geq 1$, $\epsilon > 0$ and compact $K \subset \mathcal{X}$.

New Approach via *V*-uniform Ergodicity: Assumptions

(1) $\{x_t\}$ is aperiodic, irreducible (w.r.t. measure φ) and has stationary distribution π^*.

(2) $\exists$ measurable $V : \mathcal{X} \to [1, \infty)$ s.t. $\int V(x) d\pi^*(x) < \infty$, and as $n \to \infty$,

$$\sup_{x \in \mathcal{X}, |g| \leq V} \{|E_x g(x_n) - \int g(x) d\pi^*(x)|/V(x)\} \to 0.$$

Condition (2) is called *V-uniform ergodicity (strong stability)* and is equivalent, under (1), to:

(2′) $\exists V : \mathcal{X} \to [1, \infty)$, measurable set C and constants $b, \beta > 0$ s.t. $E_x V(x_1) - V(x) \leq -\beta V(x) + b\mathbf{1}_C(x), \forall x \in \mathcal{X}$.

- V is a *modified Foster–Liapounov* function.

- Condition $(2')$ is satisfied by usual time series models and stochastic systems in signal processing, control engineering (*Meyn and Tweedie, 1993*).

- Condition (2), or its equivalent $(2')$, implies that the weak convergence of x_n to the stationary distribution π^* is geometrically fast, i.e., $\exists\ 0 < \rho < 1$ s.t. for all large n,

$$\sup_{x \in \mathcal{X}, |h| \leq V} \left\{ |E_x h(x_n) - \int h(x) d\pi^*(x)| / V(x) \right\} \leq \rho^n.$$

Goal: A similar result for π_n.

Stationary Distribution of Filter

- Suppose x_0 is initialized at π^*. Then

$$(x_t, y_t, y_{t-1}, \ldots, y_1) \underset{=}{\mathcal{L}} (x_0, y_0, y_{-1}, \ldots, y_{-t+1}).$$

- Lévy's theorem:

$$\mathcal{L}(x_0 | y_0, y_{-1}, \ldots, y_{-t+1}) \to \mathcal{L}(x_0 | y_0, y_{-1}, \ldots) \text{ a.s.}$$

- Ψ : distribution of random measure $\mathcal{L}(x_0 | y_0, y_{-1}, \ldots)$.
- $(x_t, \pi_t) \Rightarrow (\pi^*, \Psi)$ exponentially fast?

Theorem.

Let $0 < \alpha < 1$. *Under* (1) *and* (2), *if*

$$\int V(x) d\nu_0(x) + \int V(x) d\pi_0(x) < \infty,$$

then there exists $0 < \rho < 1$ *s.t. as* $n \to \infty$,

$$\sup_{|h| \leq V^\alpha} E_{\nu_0} \left| \int h(x) d\pi_n(x) - E_{\nu_0}[h(X_n)|Y_1, \ldots, Y_n] \right| = \mathcal{O}(\rho^n),$$

$$\sup_{|h| \leq V^\alpha} E_{\pi^*} \left| \int h(x) d\pi_n(x) - E_{\pi^*}[h(X_n)|Y_n, Y_{n-1}, \ldots] \right| = \mathcal{O}(\rho^n).$$

Remark.

Let π_n^0 denote the HMM filter initialized at the true initial distribution ν_0. Previous results aim at bounding $\|\pi_n - \pi_n^0\|_{TV}$ under uniform recurrence assumption $\lambda_* \leq p(x, x') \leq \lambda^*$ and thereby showing that

$$\limsup n^{-1} \log \|\pi_n - \pi_n^0\|_{TV} \leq -\lambda_*/\lambda^* \quad \text{a.s.}$$

Note that under P_{ν_0},

$$\int h(x) d\pi_n(x) - E_{\nu_0}[h(X_n)|Y_1, \ldots, Y_n] = \int h \; d(\pi_n - \pi_n^0).$$

Instead of bounded h for TV norm, we allow h to be unbounded but to grow no larger than V^α. To extend to more general chains, our approach relies on averaging $|\int h d(\pi_n - \pi_n^0)|$ under P_{ν_0} and using the tower property of conditional expectations for E_{ν_0}.

Proof.

(1) *Key Representation*. For given observations $y_1, \ldots, y_n$,

$$\int h \, d\pi_n - \int h \, d\pi_n^0$$

$$= E[h(X_n')|Y_1 = y_1, \ldots, Y_n = y_n] - E[h(X_n^*)|Y_1 = y_1, \ldots, Y_n = y_n]$$

$$= \frac{E\{[h(X_n') - h(X_n^*)] \prod_{i=1}^n g(y_i|X_i')g(y_i|X_i^*)\}}{E\{\prod_{i=1}^n g(y_i|X_i')g(y_i|X_i^*)\}},$$

where $\{X_t^*\}$ and $\{X_t'\}$ are independent Markov chains having the same transition density function $p(\cdot, \cdot)$ and such that X_0^* has distribution ν_0, X_n' has distribution π_0.

(2) Use coupling argument (via Nummelin's splitting technique) for the independent chains X_t' and X_t^* to show that

$$E\{\mathbf{1}_{\{\tau<n\}}[h(X_n') - h(X_n^*)] \prod_{i=1}^{n} g(y_i|X_i')g(y_i|X_i^*)\} = 0,$$

where τ is the coupling time of the chains X_n' and X_n^*. Put independent $\{X_t'\}, \{X_t^*\}, \{(X_t, Y_t)\}$ on same probability space and let

$$L_n = \Big\{ \prod_{i=1}^{n} g(Y_i|X_i')g(Y_i|X_i^*) \Big\} \Big/ [f_{\pi_0}(Y_1, \ldots, Y_n) f_{\nu_0}(Y_1, \ldots, Y_n)],$$

where f_π is the density of $(Y_1, \ldots, Y_n)$ when $X_0 \sim \pi$.
It remains to show

$$\sup_{|h| \le V^\alpha} E\{E[(|h(X_n')| + |h(X_n^*)|)L_n \mathbf{1}_{\{\tau \ge n\}} | Y_1, \ldots, Y_n]\} = \mathcal{O}(\rho^n).$$

(3) Use tower property $E\{E(\cdot|Y_1,\ldots,Y_n)\} = E(\cdot)$ to express the preceding expectation as

$$E\{L_n[\,|h(X'_n)| + |h(X^*_n)|\,]\mathbf{1}_{\{\tau \geq n\}}\}.$$

- The coupling time has an exponentially small tail probability.

- In the product space, L_n is the likelihood ratio for an alternative measure Q under which $(X'_1,\ldots,X'_n)$ and $(X^*_1,\ldots,X^*_n)$ are conditionally independent given $(Y_1,\ldots,Y_n) = (y_1,\ldots,y_n)$, with respective conditional densities

$$\int \prod_{i=1}^{n} p(x'_{i-1},x'_i)g(y_i|x'_i)d\pi_0(x'_i)/f_{\pi_0}(y_1,\ldots,y_n),$$

$$\int \ldots\, d\nu_0(x^*_i)/f_{\nu_0}(\ldots);$$

hence $L_n = dQ_n/dP_n$ is exponentially small under P.

- $E\{L_n[\,|h(X'_n)| + |h(X^*_n)|\,]\mathbf{1}_{\{\tau \geq n\}}\} =$
  $E\{L_n[\,|h(X'_n)| + |h(X^*_n)|\,]|\tau \geq n\}P\{\tau \geq n\}.$

Applications of Asymptotic Stability

(1) Parameter estimation in HMM:

- Block likelihood
- LAN: Let $\boldsymbol{y}_{i,j} = (y_i, \ldots, y_j)$ for $i \leq j$.
- Important tool for analysis of $D^2 \log f_\theta(y_0|\boldsymbol{y}_{-k,-1})$

$$
\begin{aligned}
&= E_\pi\{D^2 \log f_\theta(\boldsymbol{x}_{-k,0}, \boldsymbol{y}_{-k,0})|\boldsymbol{y}_{-k,0}\} \\
&\qquad - E_\pi\{D^2 \log f_\theta(\boldsymbol{x}_{-k,-1}, \boldsymbol{y}_{-k,-1})|\boldsymbol{y}_{-k,-1}\} \\
&+ \mathrm{Cov}_\pi\{D \log f_\theta(\boldsymbol{x}_{-k,0}, \boldsymbol{y}_{-k,0})|\boldsymbol{y}_{-k,0}\} \\
&\qquad - \mathrm{Cov}_\pi\{D \log f_\theta(\boldsymbol{x}_{-k,-1}, \boldsymbol{y}_{-k,-1})|\boldsymbol{y}_{-k,-1}\}
\end{aligned}
$$

is to use asymptotic stability of HMM filter which is combined with that of the time-reversed filter via Bayes' theorem to yield

$$E_{\pi^*}|E\{h(y_0, x_0, x_1)|\boldsymbol{y}_{-n,m}, x_{-n}\} - E_{\pi^*}\{h(y_0, x_0, x_1)|\boldsymbol{y}_{-n,m}\}| = \mathcal{O}(\rho^n),$$

showing that $\{D^2 \log f_{\theta_0}(y_0|\boldsymbol{y}_{-k,-1}),\ k \geq 1\}$ is Cauchy in L_1.

- To show the sequence is uniformly Cauchy for $\|\theta - \theta_0\| < \delta$, change measure under P_θ to that under P_{θ_0}.
  - Bound likelihood ratio (which is a Markov random walk) by exploiting that $\|\theta - \theta_0\| < \delta$ is sufficiently small and using large deviation bounds.
- Previous results on LAN and asymptotic efficiency require finite state spaces: *Bickel, Ritov and Rydén*, or restrictive assumptions for compact state spaces: *Jensen and Petersen*.

(2) Approximate $\mathcal{L}(X_n|Y_1, \ldots, Y_n)$ by $\mathcal{L}(X_n|Y_n, \ldots, Y_{n-m})$:

- Bounded complexity via sliding windows

(3) Change-point models:

- AR models with piecewise constant regression parameters and error variances: *Lai, Liu and Xing (2005)*
- Adaptive control of stochastic systems with time-varying parameters: *Chen and Lai (2007)*

(4) Sequential Monte Carlo implementation:

- Resampling step does not result in serious propagation of errors because of exponential forgetting factor