Nonlinear Filtering and Systems Theory

Ramon van Handel

***Abstract*— The fundamental connection between the stability of linear filtering and linear systems theory was already remarked in Kalman's seminal 1960 paper. Unfortunately, the linear theory relies heavily on the investigation of the explicit Kalman filtering equations, and sheds little light on the behavior of nonlinear filters. Nonetheless, it is possible to establish surprisingly general connections between the stability of nonlinear filters and nonlinear counterparts of basic concepts in linear systems theory: stability, observability, detectability. The proofs of these results are probabilistic in nature and provide significant insight into the mechanisms that give rise to filter stability. The aim of this paper is to review these recent results and to discuss some of their applications.**

I. Introduction

A *hidden Markov model* is defined by a Markov chain $(X_k)_{k\geq 0}$, together with an observation process $(Y_k)_{k\geq 1}$ which is conditionally independent given $(X_k)_{k\geq 0}$ with the conditional law of Y_k depending on X_k only. The dependence structure of this process is visualized in Figure 1. Think of $(X_k)_{k\geq 0}$ as the time evolution of a quantity of interest, which is not directly observable. Instead, at each time k an observation Y_k is made available to us, which is a noisy function on the current state X_k of the hidden process. A bivariate stochastic process (X,Y) of this form is perhaps the quintessential model of a partially observed system, and such models therefore appear in a wide variety of applications.[1]

As the process of interest $(X_k)_{k\geq 0}$ is not directly observable, it is a basic problem to estimate it from the observed data. To this end, we investigate the *nonlinear filter*

$$\pi_k = \mathbf{P}[X_k \in \cdot\,|Y_1,\ldots,Y_k],$$

which is simply the conditional distribution of the hidden process given the observation history. In principle, computing the filter solves the problem of estimating the hidden process optimally (in the mean square sense). On the other hand, even disregarding computational issues, it should be noted that the filter depends on our knowledge of the law $\mathbf{P}$, which must be estimated in practice by some statistical procedure. It is not at all clear how robust the filter is to model misspecification. This is particularly worrisome with regard to the initial measure (the law of X_0) which cannot be estimated at all from a single sequence of observations, as it is sampled only once in generating the time series. Therefore, in order for the filter to be of any practical use in tracking the hidden process, it must be robust to substantial misspecification of the initial measure. To make this more precise, denote by $\mathbf{P}^\mu$ the law of the hidden Markov model with initial measure $X_0 \sim \mu$, and denote by π_k^μ the filter computed using this measure. The desired property of the filter can now be stated as follows: any useful filter will be *stable* in the sense that

$$\|\pi_k^\mu - \pi_k^\nu\| \xrightarrow{k\to\infty} 0$$

for a large class of measures (μ,ν) and for a suitable norm $\|\cdot\|$ and mode of convergence. Here I am being intentionally vague, as the details will differ depending on the setting.

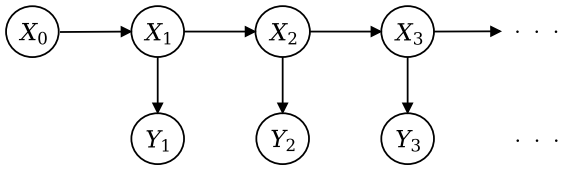

Fig. 1. The dependence structure of a hidden Markov model.

From the point of view of nonlinear filtering, the simplest hidden Markov model is the linear-Gaussian model

$$X_{k+1} = AX_k + B\xi_{k+1}, \tag{1}$$
$$Y_k = CX_k + \eta_k, \tag{2}$$

where A,B,C are matrices of appropriate dimensions and $(\xi_k,\eta_k)_{k\geq 1}$ are i.i.d. Gaussian random vectors. When in addition X_0 is Gaussian, then the filter π_k is a random Gaussian measure for every k, whose mean and covariance satisfy the celebrated Kalman filtering equations. Already in Kalman's seminal 1960 paper, the filter stability problem was raised and its solution alluded to [1, Remark (i)]. It turns out that there is a deep connection beteen stability of the Kalman filter and linear systems theory: for example, the Kalman filter is stable when the linear control system

$$x(k+1) = Ax(k) + Bu(k), \tag{3}$$
$$y(k) = Cx(k) \tag{4}$$

is controllable and observable. Unfortunately, the proofs of such results rely entirely on the fact that the Kalman filter consists of a linear recursion and a Riccati equation, whose asymptotics can be studied directly. Such results therefore shed little light on filter stability in general hidden Markov models, and seem (to this author) even somewhat mysterious in the linear-Gaussian case as the proofs are not probabilistic.

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. rvan@princeton.edu

[1]Note that we will always work in general (Polish) state spaces unless stated otherwise. Some authors use the term *hidden Markov model* exclusively for the case where X_k (and perhaps Y_k) takes values in a finite state space. However, such a definition is unnecessarily restrictive. Our setting encompasses a large number of models that appear in the literature under different names: finite state hidden Markov models, state space models, linear-Gaussian models, etc. There is also a natural generalization to continuous time, known as *Markov additive processes*. Almost every result in this paper has an equivalent continuous time counterpart, but we restrict our discussion to the discrete time setting for simplicity.

ISBN 978-963-311-370-7

In the nonlinear case, the filter stability property takes on added significance. First, in almost any case of practical interest, the nonlinear filter is a truly infinite dimensional object (a random probability measure) which does not admit a finite dimensional sufficient statistic as is the case for the linear-Gaussian model. This means that substantial computational effort is involved in the approximate implementation of the nonlinear filter. If the filter is not stable, so that reliable estimates cannot be obtained, there is no point in putting in the effort. This point was raised in the 1968 book of Bucy and Joseph, who distinguish between local theory (the problem of obtaining explicit equations for π_k) and global theory (stability and asymptotic properties of π_k):

> "In the case of nonlinear filtering, much remains to be done, for as yet only a local theory of filtering has been established, and [...] the problem of synthesis of the filter is unexplored. In fact, because the relevant regularity conditions are not sufficiently well understood, even approximate synthesis techniques are fraught with difficulties. For example, suppose a non-linear model is not observable, then it is rather wasteful to perform the computations necessary to determine the conditional distribution, because any estimate, including the optimal one, will perform poorly. Hence, it is critical for applications to develop effective sufficient conditions for observability." [2, Ch. XI]

Second, it turns out that the filtering stability property is not only of intrinsic interest, but is a key technical tool in establishing other properties of the nonlinear filter which hold over the infinite time horizon. For example, using filter stability, one can establish time-uniform robustness results for much more general misspecifications of the law of the process, as well as time-uniform approximations of the filter by numerical algorithms. Another example is the characterization of vanishing stationary estimation error for nonlinear filters in the high SNR regime, which was resolved for the Kalman filter by Kwakernaak and Sivan [3].

The goal of this paper is to review a collection of recent results which establish, in a surprisingly general setting, connections between the stability of nonlinear filters and systems theory. The discussion is heavily biased towards problems I have worked on, and I make no attempt to do justice to the breadth of the literature on filter stability, which is reviewed in detail in [4]. The work in this area has been largely dominated by beautiful *quantitative* results, which however necessarily require strong assumptions on the model. Instead, we aim to impose minimal assumptions and to elucidate *qualitatively* the fundamental mechanisms that give rise to filter stability. As one might expect, the resulting theory has a distinct systems-theoretic flavor.

The remainder of this paper is organized as follows. In section II, we revisit the linear-Gaussian case and some additional examples in order to develop an intuition for the filter stability problem. Section III is concerned with the case where the hidden process is itself stable (i.e., ergodic). Sections IV and V introduce general notions of observability, controllability, and detectability, and establish their relevance to the filter stability problem. Finally, in section VI we briefly discuss two representative applications of these results.

II. Examples and intuition

A. The Kalman filter

Despite that the stability theory for the Kalman filter does not extend to the nonlinear setting, we can use it to gain some intuition about the stability property. We begin by reviewing some notions from linear systems theory.

Definition II.1. The linear control system (3)–(4) is called *asymptotically stable* if $\|x(k)-x'(k)\| \to 0$ as $k \to \infty$ for any x,x' (and any control $u(k)$), where $x(k)$ and $x'(k)$ satisfy the recursion (3) with initial conditions $x(0)=x$ and $x'(0)=x'$.

Definition II.2. The linear control system (3)–(4) is called *observable* if there exist no $x \neq x'$ such that the initial conditions $x(0)=x$ and $x(0)=x'$ give rise to the same observation sequence $(y(k))_{k\geq 1}$ (for any control $u(k)$).

Definition II.3. The linear control system (3)–(4) is called *detectable* if for any x,x', one of the following holds:

1) either $x(0)=x$ and $x(0)=x'$ give rise to distinct observation sequences $(y(k))_{k\geq 1}$; or
2) $\|x(k)-x'(k)\| \to 0$ as $k \to \infty$, where $x(k)$ and $x'(k)$ are defined as in Definition II.1.

Definition II.4. The linear control system (3)–(4) is called *controllable* if for any x,x', there exists $n \geq 1$ and a control $(u(k))_{k<n}$ such that $x(0)=x$ and $x(n)=x'$.

Definition II.5. The linear control system (3)–(4) is called *stabilizable* if for any x,x', one of the following holds:

1) either there exists $n \geq 1$ and a control $(u(k))_{k<n}$ such that $x(0)=x$ and $x(n)=x'$; or
2) $\|x(k)-x'(k)\| \to 0$ as $k \to \infty$, where $x(k)$ and $x'(k)$ are defined as in Definition II.1.

We can now formulate a standard result on stability of the Kalman filter; see, e.g., [5] (in continuous time).

Theorem II.6. *Suppose that the linear control system (3)–(4) is detectable and stabilizable. Then the Kalman filter associated to the hidden Markov model (1)–(2) is stable in the sense that* $\mathbf{E}\|m_k - m_k'\| \to 0$ *and* $\|\Sigma_k - \Sigma_k'\| \to 0$ *as* $k \to \infty$ *for any* $m_0, m_0', \Sigma_0, \Sigma_0'$. *Here* m_k *and* Σ_k *are the mean and covariance of the random Gaussian measure* π_k^μ *with* $\mu \sim N(m_0,\Sigma_0)$, *and* m_k', Σ_k' *are defined similarly.*

One can make various improvements to this theorem by considering non-Gaussian initial measures or different notions of convergence (see [5]). The key point for us, however, is that stability is guaranteed by the systems-theoretic properties of detectability and stabilizability. Let us cite one more result [6], which shows that stabilizability is not of essence if we are willing to slightly restrict the class of initial measures. This will help us understand the distinct roles of detectability and stabilizability in Theorem II.6.

Theorem II.7. *Suppose that the linear control system (3)–(4) is detectable. Then the Kalman filter associated to the hidden Markov model (1)–(2) is stable in the sense of Theorem II.6 for any m_0, m_0' and $\Sigma_0 > 0$, $\Sigma_0' > 0$.*

B. Some intuition

Theorems II.6 and II.7 suggest that detectability is the key property that yields stability of the Kalman filter. The detectability property interpolates between two extreme cases: if the model is either asymptotically stable or observable, it is detectable. Each of these cases has an intuitive interpretation.

- The filter stability property states that the optimal estimate of the hidden process given the observations "forgets" the initial measure over a long time horizon. On the other hand, the asymptotic stability property states that the hidden process itself "forgets" its initial condition over a long time horizon. If the hidden process does not depend on the initial measure, then (intuitively) neither should its optimal estimate, so filter stability should follow from asymptotic stability.
- The observability property states that the observations as so "informative" that they essentially reveal the initial condition of the hidden process. Even if we misspecify the initial measure, the information contained in the observations will eventually obsolete the prior information contained in the initial measure. Therefore, (intuitively) the optimal estimate of the hidden process given the observations should not depend on the initial measure, so filter stability should follow from observability.

The detectability property can now be understood as the synthesis of these two mechanisms: roughly speaking, a detectable model is a model that can be split into an asymptotically stable part and an observable part.

Of course, the above discussion is merely a fanciful interpretation of the precise result obtained by studying the Kalman filtering equations. The challenge we now face is to turn these naive and somewhat vague intuitions into rigorous mathematics, in the setting of general hidden Markov models. The proofs of the Kalman filter results are of essentially no use here, so we must start from scratch.

We will tackle our goal piece by piece. In section III, we will consider the case where the hidden process forgets its initial condition. In section IV, we will consider the case of informative observations. This will give natural counterparts of the asymptotic stability and observability properties for general hidden Markov models. We will also address further the role of controllability and stabilizability. Finally, in section V we generalize the notion of detectability. Here, however, a general result is still out of reach, but we can obtain a complete characterization in the important special case where the hidden process takes values in a finite set.

C. A counterexample

To highlight the fact that the intuition developed in the previous section should not be taken for granted, let us briefly discuss a vexing example which has appeared in various guises in the literature (e.g., [7]). Let the hidden process be a finite state Markov chain in the state space $E = \{0,1,2,3\}$ with transition probability matrix

$$P = \begin{pmatrix} 1-p & p & 0 & 0 \\ 0 & 1-p & p & 0 \\ 0 & 0 & 1-p & p \\ p & 0 & 0 & 1-p \end{pmatrix}$$

for some $p > 0$. We define the observations process as the nonrandom function $Y_k = \mathbf{1}_{\{1,3\}}(X_k)$. This is a perfectly respectable hidden Markov model. At each time, the hidden process jumps to the next state (modulo 4) with probability p, and stays put otherwise. The observations tell us precisely whether we are in one of the subsets $\{0,2\}$ or $\{1,3\}$, but we cannot distinguish what state we are in within each set.

It is a straightforward exercise to compute explicitly the behavior of the nonlinear filter $\pi_k^\mu(i) = \mathbf{P}^\mu[X_k = i | Y_1, \ldots, Y_k]$. Let us therefore simply describe the conclusion. Suppose it is revealed that $Y_1 = 0$. Then we know that $X_1 \in \{0,2\}$, so the filter π_1^μ will put some mass a on the point 0 and the remaining mass $b = 1-a$ on the point 2. From this point onwards, the observations process $(Y_k)_{k\geq 1}$ reveals exactly when the hidden process transitions to the next state, but no further information is revealed that will help us distinguish between the points $0,2$ (or $1,3$, depending on the value of the current observation). Thus π_k^μ is obtained from π_1^μ by rotating the distribution at those times when the observation process reveals a jump has occured. The following table illustrates this behavior along one potential observation path.

k	1	2	3	4	5	6	7	8	9
Y_k	0	0	1	0	0	0	1	1	0
$\pi_k(0)$	a	a	0	b	b	b	0	0	a
$\pi_k(1)$	0	0	a	0	0	0	b	b	0
$\pi_k(2)$	b	b	0	a	a	a	0	0	b
$\pi_k(3)$	0	0	b	0	0	0	a	a	0

If we change the initial measure from μ to ν, we obtain exactly the same result but with different constants a, b. It is therefore immediately clear that $\|\pi_k^\mu - \pi_k^\nu\|$ is positive and constant, so the filter is not stable. On the other hand, the hidden process $(X_k)_{k\geq 0}$ is asymptotically stable in a very strong sense: it is uniformly geometrically ergodic, i.e.,

$$\sup_{\mu,\nu,i} |\mathbf{P}^\mu[X_k = i] - \mathbf{P}^\nu[X_k = i]| \xrightarrow{k\to\infty} 0$$

at a geometric rate. So we evidently have asymptotic stability of the hidden process while filter stability fails.

In view of this example, one might fear that the intuition gained from the Kalman filter does not apply in the nonlinear case and must be abandoned. Fortunately, it turns out that counterexamples of this type are extemely fragile and disappear if we impose mild nondegeneracy assumptions. For example, if we add arbitrarily small noise to the observations, e.g., $Y_k = \mathbf{1}_{\{1,3\}}(X_k) + \eta_k$ where $(\eta_k)_{k\geq 1}$ are i.i.d. $N(0,\varepsilon)$ with $\varepsilon > 0$, the filter will be stable. Alternatively, if we slightly perturb one of the transition probabilities in our example, the filter will be stable due to observability (our original example can clearly not be observable in any reasonable sense due

to the symmetry in the model). We will see in the following sections that our intuition holds true in a surprisingly general setting, but the above example warns us that we must be very careful in formulating the appropriate assumptions.

III. The ergodic case

A. A general result

The goal of this section is to make precise the following intuition: if the hidden process "forgets" its initial condition, then so does the filter. The requisite stability property of the hidden process is made precise by the following assumption, which replaces asymptotic stability in the linear setting. Here $\|\rho-\rho'\|_{\mathrm{TV}}$ denotes the total variation distance.

Assumption III.1. There is a probability measure λ so that $\|\mathbf{P}^\mu[X_k\in\cdot]-\lambda\|_{\mathrm{TV}}\to 0$ as $k\to\infty$ for any initial measure μ.

If Assumption III.1 holds, the hidden process is said to be *ergodic*. More precisely, it is well known in the theory of Markov chains that Assumption III.1 holds if and only if the hidden process is positive Harris recurrent and aperiodic.

As is demonstrated by the example in section II-C, Assumption III.1 is not quite enough to guarantee stability of the filter. We need a mild *nondegeneracy* assumption.

Assumption III.2. There is a strictly positive function $g(x,y)>0$ and a measure $\varphi(dy)$ such that $\mathbf{P}^\mu[Y_k\in A|X_k]=\int_A g(X_k,y)\,\varphi(dy)$ for every k,A,μ.

Assumption III.2 states that the conditional law of the observations possesses a positive density g (with respect to some reference measure φ). It is essentially equivalent to the following statement: for every x,x', the conditional laws $\mathbf{P}^\mu[Y_k\in\cdot\,|X_k=x]$ and $\mathbf{P}^\mu[Y_k\in\cdot\,|X_k=x']$ are absolutely continuous. Thus observation of Y_k cannot give us any *precise* information on X_k, which rules out the example in section II-C. On the other hand, an arbitrarily small amount of noise will immediately force Assumption III.2 to hold.

Our general result is now easily stated [8, Corollary 5.5].

Theorem III.3. *Suppose that Assumptions III.1 and III.2 hold. Then the nonlinear filter is stable in the sense that* $\|\pi_k^\mu-\pi_k^\nu\|_{\mathrm{TV}}\to 0$ $\mathbf{P}^\gamma$*-a.s. as* $k\to\infty$ *for any* μ,ν,γ.

The assumptions of this result are intuitive, but its proof is long and fairly technical. Nonetheless, the idea behind the proof is not difficult to understand, and provides significant insight into how filter stability is inherited from ergodicity and what goes wrong in the absence of nondegeneracy.

B. Orey's theorem

To understand the proof of Theorem III.3, we must recall a classic result in the general theory of Markov chains.

Theorem III.4 (Orey's theorem)**.** *Assumption III.1 holds if and only if the Markov chain* $(X_k)_{k\ge0}$ *possesses an invariant probability measure* λ*, and the tail* σ*-field*

$$\mathscr{T}=\bigcap_{m\ge0}\sigma\{X_k:k\ge m\}$$

is $\mathbf{P}^\mu$*-trivial for every initial measure* μ.

A complete proof can be found in [9, Ch. 6, Theorem 1.8]. However, to understand Theorem III.3, it will be helpful to give a different proof of the sufficiency part of the result.

Proof of sufficiency. Let $\mu\ll\nu$ be absolutely continuous initial measures. Then $d\mathbf{P}^\mu/d\mathbf{P}^\nu=(d\mu/d\nu)(X_0)$, so

$$\frac{d\mathbf{P}^\mu[X_k\in\cdot]}{d\mathbf{P}^\nu[X_k\in\cdot]}=\mathbf{E}^\nu\Big[\frac{d\mu}{d\nu}(X_0)\Big|X_k\Big].$$

As $\|\rho-\rho'\|_{\mathrm{TV}}=\int|\frac{d\rho}{d\rho'}-1|\,d\rho'$, we get

$$\|\mathbf{P}^\mu[X_k\in\cdot]-\mathbf{P}^\nu[X_k\in\cdot]\|_{\mathrm{TV}}=\mathbf{E}^\nu\Big[\Big|\mathbf{E}^\nu\Big[\frac{d\mu}{d\nu}(X_0)\Big|X_k\Big]-1\Big|\Big].$$

But the Markov property states that the past and future are conditionally independent given the present. In particular, X_0 and $\sigma\{X_r:r>k\}$ are conditionally independent given X_k:

$$\mathbf{E}^\nu\Big[\frac{d\mu}{d\nu}(X_0)\Big|X_k\Big]=\mathbf{E}^\nu\Big[\frac{d\mu}{d\nu}(X_0)\Big|\sigma\{X_r:r\ge k\}\Big].$$

Therefore, the martingale convergence theorem gives

$$\lim_{k\to\infty}\|\mathbf{P}^\mu[X_k\in\cdot]-\mathbf{P}^\nu[X_k\in\cdot]\|=\mathbf{E}^\nu\Big[\Big|\mathbf{E}^\nu\Big[\frac{d\mu}{d\nu}(X_0)\Big|\mathscr{T}\Big]-1\Big|\Big].$$

But $\mathscr{T}$ is $\mathbf{P}^\nu$-trivial, so $\mathbf{P}^\nu[\frac{d\mu}{d\nu}(X_0)|\mathscr{T}]=\mathbf{P}^\nu[\frac{d\mu}{d\nu}(X_0)]=1$.

We have shown that $\|\mathbf{P}^\mu[X_k\in\cdot]-\mathbf{P}^\nu[X_k\in\cdot]\|_{\mathrm{TV}}\to0$ as $k\to\infty$ whenever $\mu\ll\nu$. Now consider any initial measure μ, and let $\nu=(\mu+\lambda)/2$. Then $\mu\ll\nu$ and $\lambda\ll\nu$, so

$$\begin{aligned}\|\mathbf{P}^\mu[X_k\in\cdot]-\lambda\|_{\mathrm{TV}}&\le\|\mathbf{P}^\mu[X_k\in\cdot]-\mathbf{P}^\nu[X_k\in\cdot]\|_{\mathrm{TV}}\\&\quad+\|\mathbf{P}^\lambda[X_k\in\cdot]-\mathbf{P}^\nu[X_k\in\cdot]\|_{\mathrm{TV}}\xrightarrow{k\to\infty}0,\end{aligned}$$

where we have used that λ is invariant, i.e., $\mathbf{P}^\lambda[X_k\in\cdot]=\lambda$ for all k. Thus Assumption III.1 is established. □

The key idea of this proof is that if we consider absolutely continuous initial measures, one can give an explicit expression for the limiting total variation distance in terms of some tail σ-field. This idea appears to be quite fundamental and appears in various guises in the proof of Theorem III.3 and in the observability results of section IV below.

C. Key elements of the proof of Theorem III.3

1) Representation of the limit: The beginning of the proof of Theorem III.3 is very similar to the proof of Orey's theorem given above. First, a somewhat tedious technical argument shows that is suffices to prove that

$$\|\pi_k^\mu-\pi_k^\lambda\|_{\mathrm{TV}}\xrightarrow{k\to\infty}0\quad\mathbf{P}^\mu\text{-a.s. for }\mu\ll\lambda.$$

That is, the problem is reduced to the absolutely continuous case. At this point, we will obtain an explicit representation of the limit $\lim_{k\to\infty}\|\pi_k^\mu-\pi_k^\lambda\|_{\mathrm{TV}}$ in terms of tail σ-fields.

Let $\mu\ll\lambda$. The density of $(Y_1,\ldots,Y_k,X_k)$ is given by

$$\frac{d\mathbf{P}^\mu[(Y_1,\ldots,Y_k,X_k)\in\cdot]}{d\mathbf{P}^\lambda[(Y_1,\ldots,Y_k,X_k)\in\cdot]}=\mathbf{E}^\lambda\Big[\frac{d\mu}{d\lambda}(X_0)\Big|Y_1,\ldots,Y_k,X_k\Big].$$

The density of $(Y_1,\ldots,Y_k)$ is obtained similarly. Therefore, by the Bayes formula, the *conditional* density of X_k given $Y_1,\ldots,Y_k$ is obtained by dividing these two densities:

$$\frac{d\pi_k^\mu}{d\pi_k^\lambda} = \frac{\mathbf{E}^\lambda[\frac{d\mu}{d\lambda}(X_0)|Y_1,\ldots,Y_k,X_k]}{\mathbf{E}^\lambda[\frac{d\mu}{d\lambda}(X_0)|Y_1,\ldots,Y_k]}.$$

In particular, we have

$$\|\pi_k^\mu-\pi_k^\lambda\|_{\mathrm{TV}} = \int\left|\frac{d\pi_k^\mu}{d\pi_k^\lambda}-1\right|d\pi_k^\lambda = \frac{\mathbf{E}^\lambda[\Delta_k|Y_1,\ldots,Y_k]}{\mathbf{E}^\lambda[\frac{d\mu}{d\lambda}(X_0)|Y_1,\ldots,Y_k]},$$

where we have defined

$$\Delta_k = \left|\mathbf{E}^\lambda\left[\frac{d\mu}{d\lambda}(X_0)\middle|Y_1,\ldots,Y_k,X_k\right]-\mathbf{E}^\lambda\left[\frac{d\mu}{d\lambda}(X_0)\middle|Y_1,\ldots,Y_k\right]\right|.$$

But any hidden Markov model (X,Y) is itself a Markov chain, so that X_0 and $\sigma\{(X_r,Y_r):r>k\}$ are conditionally independent given (X_k,Y_k). In particular, we can write

$$\Delta_k = \left|\mathbf{E}^\lambda\left[\frac{d\mu}{d\lambda}(X_0)\middle|\mathscr{F}^Y_{1,\infty}\vee\mathscr{F}^X_{k,\infty}\right]-\mathbf{E}^\lambda\left[\frac{d\mu}{d\lambda}(X_0)\middle|\mathscr{F}^Y_{1,k}\right]\right|,$$

where we have defined the σ-fields

$$\mathscr{F}^X_{i,j}=\sigma\{X_i,\ldots,X_j\},\qquad \mathscr{F}^Y_{i,j}=\sigma\{Y_i,\ldots,Y_j\}.$$

By the martingale convergence theorem, we obtain

$$\lim_{k\to\infty}\|\pi_k^\mu-\pi_k^\lambda\|_{\mathrm{TV}} = \frac{\mathbf{E}^\lambda[\Delta_\infty|\mathscr{F}^Y_{1,\infty}]}{\mathbf{E}^\lambda[\frac{d\mu}{d\lambda}(X_0)|\mathscr{F}^Y_{1,\infty}]}\quad \mathbf{P}^\mu\text{-a.s.}$$

(as the denominator is $\mathbf{P}^\mu$-a.s. positive and $\mu\ll\lambda$), where

$$\Delta_\infty = \left|\mathbf{E}^\lambda\left[\frac{d\mu}{d\lambda}(X_0)\middle|\bigcap_{k\ge0}\mathscr{F}^Y_{1,\infty}\vee\mathscr{F}^X_{k,\infty}\right]-\mathbf{E}^\lambda\left[\frac{d\mu}{d\lambda}(X_0)\middle|\mathscr{F}^Y_{1,\infty}\right]\right|.$$

Therefore, the filter stability property is established if we can show that the following tail σ-field identity holds:

$$\boxed{\bigcap_{k\ge0}\mathscr{F}^Y_{1,\infty}\vee\mathscr{F}^X_{k,\infty}\stackrel{?}{=}\mathscr{F}^Y_{1,\infty}\quad\mathbf{P}^\lambda\text{-a.s.}} \tag{5}$$

The validity of this measure-theoretic identity turns out to be the central problem in the qualitative asymptotic theory of nonlinear filters (see, e.g., [7], [10], [8], [11]).

2) The role of ergodicity: The identity (5) may seem rather abstract to the reader unfamiliar with the ergodic theory of Markov chains. However, the identity becomes extremely intuitive when combined with Orey's theorem. To see this, suppose that Assumption III.1 holds. Note that $\mathscr{F}^Y_{1,\infty}$ does not depend on k, so it seems that one should be able to perform the following harmless computation:

$$\bigcap_{k\ge0}\mathscr{F}^Y_{1,\infty}\vee\mathscr{F}^X_{k,\infty}\stackrel{?}{=}\mathscr{F}^Y_{1,\infty}\vee\bigcap_{k\ge0}\mathscr{F}^X_{k,\infty}=\mathscr{F}^Y_{1,\infty}\vee\mathscr{T}. \tag{6}$$

By Orey's Theorem III.4, the tail σ-field $\mathscr{T}$ is $\mathbf{P}^\lambda$-trivial, so the identity (5) would follow immediately from ergodicity.

The alert reader should now become extremely nervous. After all, we have nowhere used the nondegeneracy assumption III.2. If everything above is correct, the filter should be stable whenever the hidden process is ergodic. This is clearly false: we have seen a counterexample in section II-C! The problem is that the seemingly harmless computation (6) is incorrect: the exchange of the intersection $\cap$ and supremum $\vee$ of σ-fields is not in general allowed. This subtle error[2] is present in a number of early papers on the asymptotics of nonlinear filters, starting with the classic paper of Kunita [16] on unique ergodicity of the filter (see [7] for a list of papers which rely on the incorrect argument (6)).

It is possible, however, to obtain a correct variant of (6) by using a technique due to von Weizsäcker [12]:

Lemma III.5. *The identity (5) holds if and only if the tail σ-field $\mathscr{T}$ is $\mathbf{P}^\lambda(\,\cdot\,|\mathscr{F}^Y_{1,\infty})$-trivial $\mathbf{P}^\lambda$-a.s.*

Let us reflect on what this means. By Orey's theorem, the hidden process $(X_k)_{k\ge0}$ is ergodic if and only if $\mathscr{T}$ is $\mathbf{P}^\mu$-trivial (in particular, $\mathbf{P}^\lambda$-trivial). However, for the filter to be stable, we actually need that $\mathscr{T}$ is $\mathbf{P}^\lambda(\,\cdot\,|\mathscr{F}^Y_{1,\infty})$-trivial:

> The filter is stable when the hidden process $(X_k)_{k\ge0}$ *conditioned* on the observations $(Y_k)_{k\ge1}$ is ergodic.

It is not at all clear that if the hidden process is ergodic, it will still be ergodic once we condition on the observations. In fact, in the counterexample of section II-C it is easily seen that this is not the case. For the ergodic property to be inherited, we need the nondegeneracy Assumption III.2.

3) Conditional ergodicity and nondegeneracy: The main part of the proof of Theorem III.3 consists in proving that $(X_k)_{k\ge0}$ conditioned on the observations $(Y_k)_{k\ge1}$ is ergodic under Assumptions III.1 and III.2. The details are too long to reproduce here, but let us sketch two essential ideas.

First, we must understand the structure of the hidden process $(X_k)_{k\ge0}$ under the conditional measure $\mathbf{P}^{\lambda|Y}:=\mathbf{P}^\lambda(\,\cdot\,|\mathscr{F}^Y_{1,\infty})$. To this end, we use again the Markov property of (X,Y): as $\{(X_r,Y_r):r<k\}$ and $\{X_{k+1},Y_r:r>k\}$ are are conditionally independent given (X_k,Y_k), we have

$$\begin{aligned}\mathbf{P}^{\lambda|Y}[X_{k+1}\in A|\mathscr{F}^X_{0,k}] &= \mathbf{P}^\lambda[X_{k+1}\in A|\mathscr{F}^Y_{1,\infty}\vee\mathscr{F}^X_{0,k}] = \\ &\mathbf{P}^\lambda[X_{k+1}\in A|\mathscr{F}^Y_{k,\infty}\vee\sigma\{X_k\}] = \mathbf{P}^{\lambda|Y}[X_{k+1}\in A|X_k].\end{aligned}$$

Therefore, under the conditional measure $\mathbf{P}^{\lambda|Y}$, the hidden process $(X_k)_{k\ge0}$ is still an (albeit time-inhomogeneous) Markov chain whose time k transition probability $P^Y_k(x,A):=\mathbf{P}^\lambda[X_{k+1}\in A|\mathscr{F}^Y_{k,\infty},\ X_k=x]$ depends on the observation path $(Y_r)_{r\ge k}$. Now, in general, the ergodic theory of time-inhomogeneous Markov chains is extremely difficult. Here, however, we are in a very fortunate situation: even though the conditional Markov chain is time-inhomogeneous for each observation path individually, it is not difficult to see that the transition probabilities themselves form a *stationary* stochastic process $k\mapsto P^Y_k$ under $\mathbf{P}^\lambda$. That is, the conditioned hidden process is a *Markov chain in a random environment* in the sense of Cogburn [17]. Because of their inherent stationarity, one may develop an ergodic theory for this class

[2] The problem of exchanging the intersection and supremum of σ-fields plays an important role in many problems in different areas of probability [12], and appears to be one of its worst pitfalls [13, p. 30]. The incorrect argument (6) can be found even in the work of Kolmogorov [14, p. 837], Kallianpur and Wiener [15, pp. 91–93], and Kunita [7, pp. 649–650].

of Markov processes with random transition probabilities which is surprisingly similar to the ergodic theory of ordinary time-homogeneous chains (see [17] for a countable state space, and [8] for general state spaces). In particular, it turns out that such random Markov chains are ergodic if and only if their transition probabilities are almost surely irreducible and aperiodic, in a suitable generalized sense.

The problem therefore reduces to showing that the conditional transition probabilities P_k^Y of the hidden process are almost surely irreducible aperiodic. However, Assumption III.1 only implies that the original transition probability $P(x,A) = \mathbf{P}^\lambda[X_k \in A | X_{k-1} = x]$ is irreducible aperiodic. The essential idea that allows to complete the proof is that under the nondegeneracy Assumption III.2, one can show by means of a coupling argument that the original and conditional transition probabilities are almost surely mutually absolutely continuous $P \sim P_k^Y$, so that the irreducible aperiodic property of P is inherited by P_k^Y. [Intuitively one may think of finite state Markov chains: if two chains have mutually absolutely continuous transition probabilities, they have the same transition graph. Thus if one is ergodic, so is the other.] We therefore see that the nondegeneracy property provides the "glue" which links the ergodic theory of the conditioned chain to the ergodic theory of the original chain: in the absence of nondegeneracy this crucial connection is lost, which leads to counterexamples like the one in section II-C.

IV. Observability and controllability

A. A general result

In section II-B, we discussed two mechanisms that are expected to contribute to filter stability: asymptotic stability of the hidden process, and observability of the model. The former was developed in a general setting in the previous section. The goal of this section is to investigate the latter property in the general setting. That is, we aim to make precise the following intuition: if the observations are sufficiently informative, then the filter should "forget" its (obsolete) initial condition. The requisite notion of observability, which replaces the observability property in the linear setting, is made precise by the following definition.

Definition IV.1. A hidden Markov model is called *uniformly observable* if for every $\varepsilon > 0$, there is a $\delta > 0$ such that

$$\|\mathbf{P}^\mu|_{(Y_k)_{k\ge1}} - \mathbf{P}^\nu|_{(Y_k)_{k\ge1}}\|_{\mathrm{TV}} < \delta \quad \text{implies} \quad \|\mu-\nu\|_{\mathrm{BL}} < \varepsilon.$$

It is called *observable* if $\mathbf{P}^\mu|_{(Y_k)_{k\ge1}} = \mathbf{P}^\nu|_{(Y_k)_{k\ge1}}$ implies $\mu = \nu$.

Here we write $\mathbf{P}^\rho|_{(Y_k)_{k\ge1}} = \mathbf{P}^\rho[(Y_k)_{k\ge1} \in \cdot\,]$, and we have defined the dual bounded-Lipschitz distance as $\|\rho-\rho'\|_{\mathrm{BL}} = \sup\{|\rho(f)-\rho'(f)| : |f(x)| \le 1,\ |f(x)-f(y)| \le |x-y|\ \forall x,y\}$. In words, a hidden Markov model is called observable if distinct initial measures give rise to distinct laws of the observation process $(Y_k)_{k\ge1}$. Uniform observability is a more quantitative version of this idea: roughly speaking, the model is uniformly observable if two initial measures which give rise to nearly identical observation laws must themselves be nearly identical, in a suitable (not entirely obvious) sense.

Our general result is now as follows [18, Theorem 3.3].

Theorem IV.2. *Suppose that the hidden Markov model is uniformly observable. Then we have*

$$\|\pi_k^\mu - \pi_k^\nu\|_{\mathrm{BL}} \xrightarrow{k\to\infty} 0 \ \ \mathbf{P}^\mu\text{-a.s. when } \ \mathbf{P}^\mu|_{(Y_k)_{k\ge1}} \ll \mathbf{P}^\nu|_{(Y_k)_{k\ge1}}.$$

If (X,Y) is Feller and if X takes values in a compact state space, it suffices to assume the model is observable.

We immediately note two things. On the one hand, this result is very general: even nondegeneracy of the observations (Assumption III.2) is not required. On the other hand, the notion of stability guaranteed by this result is weaker than that of Theorem III.3: convergence holds only in the dual bounded-Lipschitz norm (rather than the stronger total variation norm), and stability is only guaranteed for initial measures with absolutely continuous observation laws. One can easily find a variety of counterexamples which show that one cannot strengthen the conclusion of the theorem without additional assumptions. For example, the model where $X_k = X_0 \in [0,1]$ for all k and $Y_k = X_k + \eta_k$ with $(\eta_k)_{k\ge1}$ i.i.d. $N(0,1)$ is clearly observable and Feller, so that Theorem IV.2 applies. However, if we choose $\mu = \delta_0$ and $\nu = \delta_1$, then $\pi_k^\mu = \delta_0$ and $\pi_k^\nu = \delta_1$ for all k, so the filter is not stable. Of course, the problem is that in this case the observation laws $\mathbf{P}^\mu|_{(Y_k)_{k\ge1}}$ and $\mathbf{P}^\nu|_{(Y_k)_{k\ge1}}$ are mutually singular. If we consider exactly the same example, except that we choose the modified dynamics $X_k = X_{k-1}/2$ for all k, we find that $\mathbf{P}^\mu|_{(Y_k)_{k\ge1}} \ll \mathbf{P}^\nu|_{(Y_k)_{k\ge1}}$ and indeed $\|\pi_k^\mu - \pi_k^\nu\|_{\mathrm{BL}} \to 0$ as is claimed by Theorem IV.2, but $\|\pi_k^\mu - \pi_k^\nu\|_{\mathrm{TV}} = 2$ for all k.

The proof of Theorem IV.2 is surprisingly easy (we give almost the entire proof in section IV-B below). A much more difficult problem is to verify whether the uniform observability assumption holds in a given model (this is in contrast to the assumptions of Theorem III.3: the verication of Assumption III.1 is a well studied problem, see [19], while Assumption III.2 is trivial). Sufficient conditions for observability are mainly restricted to *additive noise* models

$$Y_k = h(X_k) + \eta_k,$$

where $(\eta_k)_{k\ge0}$ are i.i.d. random variables with nowhere vanishing characteristic function and h is a (not necessarily invertible) observation function. The reason is that in this case, the characteristic function of $(Y_1,\dots,Y_k)$ factors into the characteristic function of $(h(X_1),\dots,h(X_k))$ and a nowhere vanishing function, so that the model is observable if and only if the noiseless model $Y_k = h(X_k)$ is observable. When the hidden process takes values in a finite state space, one can now easily give explicit necessary and sufficient conditions for observability in terms of linear algebra [6, section 6]. On the other hand, in the linear-Gaussian case, we find that the model is uniformly observable precisely when the model is observable in the sense of linear systems theory (Definition II.2), see [18, section 3.3]. In more general models, it can be shown that (uniform) observability holds for additive noise observations whenever the observation function h possesses a (uniformly continuous) inverse, see [6, section 5.3], [18, section 3.4], [20]. Let us note that verification of *uniform* observability is not entirely straightforward even in the

simplest case, as it requires us to study the inverses of convolution operators; see [18, Appendix C] for details.

B. Proof of Theorem IV.2

We now give an essentially complete proof of the first part of Theorem IV.2. Once the correct notion of observability is formulated, the proof follows almost immediately from the following classic result of Blackwell and Dubins [21] (a special case of this result was rediscovered in connection with filter stability by Chigansky and Liptser [22]). Note the strong similarity between the proof of this result and that of Orey's theorem (see Theorem III.4 above).

Theorem IV.3 (Blackwell and Dubins)**.** *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be two probability laws of a stochastic process* $(Y_k)_{k\ge1}$*. Then*

$$\|\mathbf{P}[(Y_r)_{r>k}\in\cdot\,|Y_1,\ldots,Y_k]-\mathbf{Q}[(Y_r)_{r>k}\in\cdot\,|Y_1,\ldots,Y_k]\|_{\mathrm{TV}}\to0$$

as $k\to\infty$ $\mathbf{P}$*-a.s., provided* $\mathbf{P}\ll\mathbf{Q}$ *are absolutely continuous.*

Proof. Note that, by assumption, the density of $(Y_k)_{k\ge1}$ is $d\mathbf{P}/d\mathbf{Q}$. Therefore, the density of $(Y_1,\ldots,Y_k)$ is given by

$$\frac{d\mathbf{P}[(Y_1,\ldots,Y_k)\in\cdot\,]}{d\mathbf{Q}[(Y_1,\ldots,Y_k)\in\cdot\,]}=\mathbf{E}_{\mathbf{Q}}\Big[\frac{d\mathbf{P}}{d\mathbf{Q}}\Big|Y_1,\ldots,Y_k\Big].$$

Dividing the two densities, we obtain by the Bayes formula

$$\frac{d\mathbf{P}[(Y_r)_{r>k}\in\cdot\,|Y_1,\ldots,Y_k]}{d\mathbf{Q}[(Y_r)_{r>k}\in\cdot\,|Y_1,\ldots,Y_k]}=\frac{\frac{d\mathbf{P}}{d\mathbf{Q}}}{\mathbf{E}_{\mathbf{Q}}[\frac{d\mathbf{P}}{d\mathbf{Q}}|Y_1,\ldots,Y_k]}.$$

As $\|\rho-\rho'\|_{\mathrm{TV}}=\int|\frac{d\rho}{d\rho'}-1|\,d\rho'$, we get

$$\|\mathbf{P}[(Y_r)_{r>k}\in\cdot\,|Y_1,\ldots,Y_k]-\mathbf{Q}[(Y_r)_{r>k}\in\cdot\,|Y_1,\ldots,Y_k]\|_{\mathrm{TV}}=\frac{\mathbf{E}_{\mathbf{Q}}[\;|\frac{d\mathbf{P}}{d\mathbf{Q}}-\mathbf{E}_{\mathbf{Q}}[\frac{d\mathbf{P}}{d\mathbf{Q}}|Y_1,\ldots,Y_k]|\;|Y_1,\ldots,Y_k]}{\mathbf{E}_{\mathbf{Q}}[\frac{d\mathbf{P}}{d\mathbf{Q}}|Y_1,\ldots,Y_k]}.$$

But clearly $\mathbf{E}_{\mathbf{Q}}[d\mathbf{P}/d\mathbf{Q}|Y_1,\ldots,Y_k]\to d\mathbf{P}/d\mathbf{Q}$ as $k\to\infty$ $\mathbf{Q}$-a.s. by the martingale convergence theorem (hence $\mathbf{P}$-a.s. as $\mathbf{P}\ll\mathbf{Q}$), while $d\mathbf{P}/d\mathbf{Q}>0$ $\mathbf{P}$-a.s. The result follows.[3] □

Blackwell and Dubins' theorem can be stated as follows: predicting the future of a stochastic process given its past will "forget" the difference between any two absolutely continuous probability measures. In order to apply this result to the filter, we must understand the relation between estimating the current value of the hidden process and predicting the future values of the observation process. To this end, note that, for any function $f(y_1,\ldots,y_\ell)$, we have

$$\begin{aligned}
&\mathbf{E}^{\mu}[f(Y_{k+1},\ldots,Y_{k+\ell})|Y_1,\ldots,Y_k]\\
&\quad=\mathbf{E}^{\mu}[\;\mathbf{E}^{\mu}[f(Y_{k+1},\ldots,Y_{k+\ell})|Y_1,\ldots,Y_k,X_k]\;|Y_1,\ldots,Y_k]\\
&\quad=\mathbf{E}^{\mu}[\;\mathbf{E}^{\mu}[f(Y_{k+1},\ldots,Y_{k+\ell})|X_k]\;|Y_1,\ldots,Y_k]\\
&\quad=\mathbf{E}^{\mu}[\;\mathbf{E}^{\delta_{X_k}}[f(Y_1,\ldots,Y_\ell)]\;|Y_1,\ldots,Y_k]\\
&\quad=\mathbf{E}^{\pi_k^{\mu}}[f(Y_1,\ldots,Y_\ell)],
\end{aligned}$$

where we have used the tower property of the conditional expectation and the conditional independence structure of the hidden Markov model. We can therefore write

$$\begin{aligned}
&\|\mathbf{P}^{\mu}[(Y_r)_{r>k}\in\cdot\,|Y_1,\ldots,Y_k]-\mathbf{P}^{\nu}[(Y_r)_{r>k}\in\cdot\,|Y_1,\ldots,Y_k]\|_{\mathrm{TV}}\\
&\qquad=\|\mathbf{P}^{\pi_k^{\mu}}|_{(Y_r)_{r\ge1}}-\mathbf{P}^{\pi_k^{\nu}}|_{(Y_r)_{r\ge1}}\|_{\mathrm{TV}}.
\end{aligned}$$

Theorem IV.3 therefore shows that

$$\|\mathbf{P}^{\pi_k^{\mu}}|_{(Y_r)_{r\ge1}}-\mathbf{P}^{\pi_k^{\nu}}|_{(Y_r)_{r\ge1}}\|_{\mathrm{TV}}\xrightarrow{k\to\infty}0\quad\mathbf{P}^{\mu}\text{a.s.}$$

whenever $\mathbf{P}^{\mu}|_{(Y_k)_{k\ge1}}\ll\mathbf{P}^{\nu}|_{(Y_k)_{k\ge1}}$. But then

$$\|\pi_k^{\mu}-\pi_k^{\nu}\|_{\mathrm{BL}}\xrightarrow{k\to\infty}0\quad\mathbf{P}^{\mu}\text{a.s.}$$

follows immediately from the definition of uniform observability. Once we have seen this proof, the significance of the uniform observability assumption becomes completely transparent: the Blackwell-Dubins theorem shows that the predictor of the observations is always stable; but if the observations are sufficiently informative (in the sense of uniform observability), then two initial measures which lead to nearly equal estimates of the observations will also lead to nearly equal estimates of the hidden process. Thus filter stability obtains. In my opinion, this demystifies the connection between observability and filter stability even in the linear-Gaussian case: a proof using the Kalman filtering equations does not admit such a natural probabilistic interpretation!

We omit the second part of the proof of Theorem IV.2, which shows that observability already implies uniform observability if (X,Y) is Feller and X takes values in a compact state space. This result follows from elementary weak convergence arguments, see [18, Proposition 3.5].

C. The role of controllability

In the Kalman filter case, stability results typically require some form of controllability in addition to observability (e.g., Theorem II.6). In contrast, our general Theorem III.3 merely requires observability, but the filter stability property holds only for initial measures that give absolutely continuous observation laws. This suggests that the main role of the classical controllability assumption has nothing to do with the stability property itself, but merely forces the absolute continuity condition $\mathbf{P}^{\mu}|_{(Y_k)_{k\ge1}}\ll\mathbf{P}^{\nu}|_{(Y_k)_{k\ge1}}$ to hold for any pair of initial measures μ,ν. The modified Theorem II.7 for the Kalman filter provides further evidence for this claim. [Let us note that Theorem II.7 was proved in [6] specifically in order to gain insight into the role of controllability suggested by our general observability results; it does not appear to be known in the classical literature.]

To justify this claim, we now give a general result on the absolute continuity of the observation laws.

Lemma IV.4. *Suppose that one of the following holds:*

1) $\mu\ll\nu$*; or*
2) *Assumption III.2 holds and*

$$\|\mathbf{P}^{\mu}[X_k\in\cdot\,]-\mathbf{P}^{\nu}[X_k\in\cdot\,]\|_{\mathrm{TV}}\to0\quad\textit{as } k\to\infty\textit{; or}$$

[3] We have glossed over a technicality: if $\Delta_k\to0$ a.s., this does not automatically imply $\mathbf{E}[\Delta_k|\mathscr{F}_k]\to0$ a.s. The problem is resolved by the application of Hunt's lemma and a simple truncation argument.

3) *Assumption III.2 holds and*

$$\mathbf{P}^\mu[X_k \in \cdot\,] \ll \mathbf{P}^\nu[X_k \in \cdot\,] \quad \textit{for some } k \geq 0.$$

Then $\mathbf{P}^\mu|_{(Y_k)_{k\geq 1}} \ll \mathbf{P}^\nu|_{(Y_k)_{k\geq 1}}$.

The first case is trivial and holds regardless of any features of the problem. The second case follows from [8, Lemma 3.7]. This holds under Assumption III.1, which explains why our general result for the ergodic case, Theorem III.3, does not require any absolute continuity assumptions. Finally, the third case can be proved as in the proof of [20, Proposition 2.5]. This case corresponds precisely to controllability: the linear-Gaussian model is controllable if and only if there is a $k \geq 0$ such that $\mathbf{P}^\mu[X_k \in \cdot\,]$ has a density with respect to the Lebesgue measure for any initial measure μ. Such a connection with the controllability of deterministic control systems holds much more generally for stochastic systems, particularly in continuous time where there is a deep connection with so-called support theorems [23].

Let us note that in addition to eliminating the restriction on the initial measures in Theorem IV.2, controllability properties can often be used to strengthen the convergence in the weak $\|\cdot\|_{\mathrm{BL}}$-norm given by Theorem IV.2 to the stronger $\|\cdot\|_{\mathrm{TV}}$ convergence. See [20] for the relevant arguments.

V. Detectability

A. Finite state hidden Markov models

In the previous sections, we have seen that the two intuitive mechanisms supporting the filter stability property—asymptotic stability of the hidden process, and observability of the model—can each be made mathematically precise in a surprisingly general setting. In the case of the Kalman filter, however, we could go one step further and combine the asymptotic stability and observability properties into a single detectability property. Unfortunately, to date, a counterpart of this result is not known for general hidden Markov models. However, in the special case where the hidden process X takes values in a *finite* state space, we can obtain a complete analog of the detectability result for the Kalman filter. What is more, in this setting, the detectability property is *necessary and sufficient* for stability under suitable assumptions.

Let us begin by stating the requisite result.

Definition V.1. A hidden Markov model is called *detectable* if for any initial measures μ, ν, one of the following holds:

1) either $\mathbf{P}^\mu|_{(Y_k)_{k\geq 1}} \neq \mathbf{P}^\nu|_{(Y_k)_{k\geq 1}}$; or
2) $\|\mathbf{P}^\mu[X_k \in \cdot\,] - \mathbf{P}^\nu[X_k \in \cdot\,]\|_{\mathrm{TV}} \xrightarrow{k\to\infty} 0$.

Theorem V.2. *Suppose that the hidden process takes values in a finite state space, and that the observations are nondegenerate (Assumption III.2). The following are equivalent:*

1) *The hidden Markov model is detectable.*
2) $\|\pi_k^\mu - \pi_k^\nu\|_{\mathrm{TV}} \xrightarrow{k\to\infty} 0$ $\mathbf{P}^\mu$*-a.s. if* $\mathbf{P}^\mu|_{(Y_k)_{k\geq 1}} \ll \mathbf{P}^\nu|_{(Y_k)_{k\geq 1}}$.

It should be evident that the general detectability property of Definition V.1 is a direct counterpart to the notion of detectability in linear systems theory, Definition II.3. Intuitively, each pair of initial measures can either be distinguished through the information obtained from the observations, or otherwise the dynamics of the hidden process must asymptotically "forget" their difference. Thus the two intuitive mechanisms discussed in section II-B conspire to give rise to the filter stability property, at least in two special cases: the linear-Gaussian case and the finite state case.

The finite state case is obviously much simpler than general hidden Markov models. For example, finite state Markov chains cannot be transient or null recurrent. On the other hand, finite state Markov chains are in many ways prototypical of nonlinear Markov models, and are almost opposite in nature to linear-Gaussian models. Moreover, they play an important role in many practical applications. That a complete systems-theoretic characterization of stability is even possible—albeit in this special case—suggests that there must be something fundamental about the theory we have developed. It would very interesting to find a more general underlying result, of which Theorems V.2 and II.7 are special cases, but I do not know how to achieve this goal.

Remark V.3. It is interesting to note that in the finite state case, assuming $\mathbf{P}^\mu|_{(Y_k)_{k\geq 1}} \ll \mathbf{P}^\nu|_{(Y_k)_{k\geq 1}}$ is not a big restriction. Indeed, if we choose an initial measure ν which puts mass at every point of the (finite) state space, then $\mu \ll \nu$ for any initial measure μ. Therefore, if we use the initial measure ν to compute the filter, the latter will be asymptotically optimal regardless of the unknown true initial measure μ.

B. Idea of the proof of Theorem V.2

The proof of a result similar to Theorem V.2, but in continuous time, is given in [6, section 6.2]. The proof of Theorem V.2 is a straightfoward adaptation of the proof in [6] to the discrete time setting. We presently discuss the essential ideas behind the proof, leaving the details to the reader.

The easy half of the proof is to show that detectability is a necessary condition for stability. Indeed, suppose that the filter is stable (in the sense of Theorem V.2). Either the model is observable, in which case it is obviously detectable, or the model is not observable. In the latter case, by definition, there exist initial measures $\mu \neq \nu$ such that $\mathbf{P}^\mu|_{(Y_k)_{k\geq 1}} = \mathbf{P}^\nu|_{(Y_k)_{k\geq 1}}$. Because π_k^ν is a function of $(Y_1, \ldots, Y_k)$ only, we have

$$\mathbf{E}^\mu[\pi_k^\nu] = \mathbf{E}^\nu[\pi_k^\nu] = \mathbf{P}^\nu[X_k \in \cdot\,], \quad \mathbf{E}^\mu[\pi_k^\mu] = \mathbf{P}^\mu[X_k \in \cdot\,].$$

Note that we may assume without loss of generality that $\mu \ll \nu$ (otherwise, replace ν by $(\mu+\nu)/2$). Therefore,

$$\begin{aligned}\|\mathbf{P}^\mu[X_k \in \cdot\,] - \mathbf{P}^\nu[X_k \in \cdot\,]\|_{\mathrm{TV}} &= \|\mathbf{E}^\mu[\pi_k^\mu - \pi_k^\nu]\|_{\mathrm{TV}} \\ &\leq \mathbf{E}^\mu[\|\pi_k^\mu - \pi_k^\nu\|_{\mathrm{TV}}] \xrightarrow{k\to\infty} 0,\end{aligned}$$

where we used filter stability. Thus the model is detectable.

The difficult part of the proof of Theorem V.2 is to show that detectability is a sufficient condition for filter stability. Intuitively, we would like to split the model into two parts: an observable part and an unobservable part (which is required to be asymptotically stable by the detectability assumption). Each part can then be dealt with separately

using the appropriate result in section III or IV. The splitting of the model into observable and unobservable parts is achieved by the following refinement of Theorem IV.2.

Definition V.4. For any μ,ν, write $\mu \approx \nu$ iff $\mathbf{P}^\mu|_{(Y_k)_{k\geq 1}} = \mathbf{P}^\nu|_{(Y_k)_{k\geq 1}}$. Define the *space of observable functions* as

$$\mathscr{O} = \{f : \mu(f) = \nu(f) \text{ for all } \mu \approx \nu\}.$$

Theorem V.5. *Suppose that the hidden process takes values in a finite state space. Then* $|\pi^\mu_k(f) - \pi^\nu_k(f)| \to 0$ *as* $k \to \infty$ $\mathbf{P}^\mu$*-a.s. for all* $f \in \mathscr{O}$ *whenever* $\mu \ll \nu$.

The space $\mathscr{O}$ consists of those functions whose expectation can be determined from the law of the observations; that is, $\mathscr{O}$ is the part of the model about which the observations are informative. Theorem V.5 now states that at least the observable part of the model is always stable. The proof (in continuous time and for more general models) is given in [6], and is similar in spirit to the proof of Theorem IV.2. Note that the model is observable precisely when $\mathscr{O}$ contains every function, so the usual observability result is a special case.

The key idea of the proof of Theorem V.2 is now contained in the following lemma, which crucially uses Theorem V.5.

Lemma V.6. *Suppose that the hidden process takes values in a finite state space and that the model is detectable. Let* f *be a* k*-invariant function, i.e.,* $f(X_{nk}) = f(X_0)$ $\mathbf{P}^\nu$*-a.s. for any* n,ν*. Then* $f(X_0)$ *is* $\mathbf{P}^\nu$*-a.s.* $\mathscr{F}^Y_{1,\infty}$*-measurable for any* ν.

Proof. We first claim that $f \in \mathscr{O}$. Indeed, let $\mu \approx \nu$. Then

$$|\mu(f) - \nu(f)| = |\mathbf{E}^\mu[f(X_{nk})] - \mathbf{E}^\nu[f(X_{nk})]| \xrightarrow{n\to\infty} 0,$$

where the equality follows as f is k-invariant and the limit follows as the model is detectable. Therefore, $f \in \mathscr{O}$.

Fix ν, and let $\mu = \delta_x$ for some point x with $\nu(x) > 0$. Clearly $\mathbf{E}^\mu[Z] = \mathbf{E}^\nu[Z|X_0 = x]$ for any random variable Z. By k-invariance and the martingale convergence theorem,

$$\begin{aligned}|\pi^\mu_{nk}(f) - \pi^\nu_{nk}(f)| &= |\mathbf{E}^\mu[f(X_0)|\mathscr{F}^Y_{1,nk}] - \mathbf{E}^\nu[f(X_0)|\mathscr{F}^Y_{1,nk}]| \\ &\xrightarrow{n\to\infty} |f(x) - \mathbf{E}^\nu[f(X_0)|\mathscr{F}^Y_{1,\infty}]| \quad \mathbf{P}^\mu\text{-a.s.}\end{aligned}$$

Therefore, as $f \in \mathscr{O}$, we obtain by Theorem V.5

$$\mathbf{P}^\nu[\ \mathbf{E}^\nu[f(X_0)|\mathscr{F}^Y_{1,\infty}] = f(X_0)\ |X_0 = x] = 1$$

for all x with $\nu(x) > 0$. Multiplying by $\nu(x)$ and summing over x gives $\mathbf{E}^\nu[f(X_0)|\mathscr{F}^Y_{1,\infty}] = f(X_0)$ $\mathbf{P}^\nu$-a.s., as desired. □

How do we use this result? It is well known that the (finite) state space of the hidden process can be partitioned into a set T of transient states and a finite number $C_1,\dots,C_p$ of cyclic classes (e.g., [24]). If we assume for simplicity that the hidden process does not possess any transient states, then clearly the indicator function $\mathbf{1}_{C_i}$ of any cyclic class C_i is k-invariant, where k is the cycle length of C_i. Therefore, by Lemma V.6, *we can determine with certainty from the observations* $(Y_k)_{k\geq 1}$ *which cyclic class we started in.* The Bayes formula then allows us to condition on the cyclic class, so that we may consider only the stability problem for initial measures supported in the same cyclic class [6, Lemma 14]. But given two initial measures supported in the same cyclic class of cycle length k, the extended model $(\tilde X, \tilde Y)$ given by $\tilde X_n = (X_{(n-1)k+1},\dots,X_{nk})$, $\tilde Y_n = (Y_{(n-1)k+1},\dots,Y_{nk})$ defines an ergodic hidden Markov model. Therefore, we have used the observability Theorem V.5 and the detectabilitiy condition to reduce the filter stability problem to the ergodic case, at which point Theorem III.3 can be applied.[4]

It is interesting to note that the above proof has a completely natural interpretation in terms of the fundamental identity (5) that played such a central role in the ergodic case. A classic result of Blackwell and Freedman [25] states that for a finite state Markov chain without transient states, the tail σ-field $\mathscr{T}$ is precisely generated by the cyclic events:

$$\mathscr{T} = \bigcap_{k\geq 0} \mathscr{F}^X_{k,\infty} = \sigma\{\mathbf{1}_{C_1}(X_0),\dots,\mathbf{1}_{C_p}(X_0)\} \quad \mathbf{P}^\mu\text{-a.s.}$$

(see [26] for a more general characterization of the tail σ-field of a Markov chain). Therefore, Lemma V.6 can be restated as follows: if the model is detectable, then

$$\mathscr{T} \subset \mathscr{F}^Y_{1,\infty} \qquad \mathbf{P}^\mu\text{-a.s.}$$

Thus, if only the incorrect exchange of intersection and supremum (6) were justified, we would immediately have

$$\bigcap_{k\geq 0} \mathscr{F}^Y_{1,\infty} \vee \mathscr{F}^X_{k,\infty} \overset{?}{=} \mathscr{F}^Y_{1,\infty} \vee \mathscr{T} = \mathscr{F}^Y_{1,\infty} \quad \mathbf{P}^\mu\text{-a.s.},$$

establishing (5) and therefore filter stability! However, as we know, the exchange of intersection and supremum is *not* justified in general, and indeed Theorem V.2 only holds under the additional nondegeneracy Assumption III.2. Nonetheless, these observations might hint at one approach for extending Theorem V.2 to more general hidden Markov models.

C. The Kalman filter revisited

Beside the finite state case, it is classical that detectability implies filter stability in the linear-Gaussian case (Theorem II.6). This proof of this result, however, relies entirely on a study of the Kalman filtering equations, and is not probabilistic in nature. It is therefore natural to ask whether this result can be reproduced in a probabilistic manner by using the general theory discussed in the previous sections. To this end, we presently revisit these results and discuss how they apply to the linear-Gaussian model (1)–(2).

Let us begin with the ergodic case, Theorem III.3. The nondegeneracy Assumption III.2 is automatically satisfied due to the presence of the additive noise η_k in the observation model (2). Moreover, it is well known that if the linear control system (3)–(4) is asymptotically stable in the sense of Definition II.1, the hidden process (1) possesses a unique invariant probability measure λ such that $\mathbf{P}^\mu[X_k \in \cdot\,] \to \lambda$

[4] Note that we have assumed here that there are no transient states. However, transient states do not play an important role in the problem. Indeed, suppose the transient class T is nonempty. As X is a finite state Markov chain, $\mathbf{1}_T(X_k) \to 0$ as $k \to \infty$ $\mathbf{P}^\mu$-a.s. for any μ, so that certainly $\pi^\mu_k(\mathbf{1}_T) \to 0$ as $k \to \infty$ $\mathbf{P}^\mu$-a.s. for any μ. A somewhat tedious technical argument then shows that, for the purpose of filter stability, we may ignore the transient states alltogether (see [6, section 6.2.3] for details).

weakly as $k \to \infty$ for any initial measure μ. However, in this general setting, the convergence to the invariant measure need not hold in the total variation norm $\|\cdot\|_{\mathrm{TV}}$, so that Assumption III.1 need not hold. For example, if $X_{k+1} = X_k/2$ without any added noise ($B = 0$ in (1)), then clearly $X_k \to 0$ $\mathbf{P}^\mu$-a.s. for any initial measure μ. Therefore $\lambda = \delta_0$ and $\mathbf{P}^\mu[X_k \in \cdot\,] \to \lambda$ weakly for any μ, but $\|\mathbf{P}^\mu[X_k \in \cdot\,] - \lambda\|_{\mathrm{TV}} = 2$ for all k when μ has a density with respect to the Lebesgue measure (as then $\mathbf{P}^\mu[X_k \in \cdot\,]$ has a density for any k, and is therefore singular with respect to λ).

In order for Assumption III.1 to hold for the linear-Gaussian model, asymptotic stability is evidently not enough: we must ensure that there is enough noise in the hidden process so that convergence holds in total variation. It is well known that controllability is necessary and sufficient for the latter to be the case. In the linear-Gaussian case, our general Theorem III.3 therefore reduces to the following result:

Corollary V.7. *Suppose that the linear control system (3)–(4) is asymptotically stable and controllable. Then the filter associated to the linear-Gaussian model (1)–(2) is stable, i.e., $\|\pi_k^\mu - \pi_k^\nu\|_{\mathrm{TV}} \to 0$ $\mathbf{P}^\gamma$-a.s. as $k \to \infty$ for any μ, ν, γ.*

We now turn to the observable case, Theorem IV.2. We have already discussed in section IV that the linear-Gaussian model (1)–(2) is uniformly observable precisely when the linear control system (3)–(4) is observable in the sense of Definition II.2. Using this and Lemma IV.4, we find that our general Theorem IV.2 reduces to the following result:

Corollary V.8. *Suppose that the linear control system (3)–(4) is observable. Then the filter associated to the linear-Gaussian model (1)–(2) is stable in the sense that*

$$\|\pi_k^\mu - \pi_k^\nu\|_{\mathrm{BL}} \xrightarrow{k\to\infty} 0 \;\; \mathbf{P}^\mu\text{-a.s. when } \; \mathbf{P}^\mu|_{(Y_k)_{k\ge1}} \ll \mathbf{P}^\nu|_{(Y_k)_{k\ge1}}.$$

If, in addition, the linear control system (3)–(4) is controllable, then this stability property holds for arbitrary μ, ν.

We conclude that the general theory developed in the previous sections allows us to establish stability results for the Kalman filter, in some important special cases, using entirely probabilistic proofs. I would argue that such proofs are much more satisfying than the original proofs which rely on the Kalman filtering equations, as the probabilistic proofs are "intrinsic" to the filtering problem and really elucidate the underlying reasons for the stability of the filter. On the other hand, the assumptions of the above Corollaries fall short of the generality of Theorems II.6 and II.7 (our conclusions are also somewhat stronger, e.g., we do not require Gaussian initial measures, but this is not a major difficulty: Theorems II.6 and II.7 can be strengthened along the lines of [5]). Therefore, the Kalman filter remains an important test case for generalizations of the theory described in this paper.

VI. Two applications

The goal of this final section is to discuss some selected applications of the theory discussed in this paper. Due to space constraints, we are limited here to discussing two representative applications. Of course, these are by no means the only possible applications: indeed, the filter stability property turns out to play a direct or indirect role in almost every problem in which the nonlinear filter is of interest on the infinite time horizon. Despite the brevity of this section, I hope to convince the reader by means of these two examples that stability properties of nonlinear filters are not only of interest in their own right, but play a role in quite disparate problems involving hidden Markov models.

A. The maximal accuracy problem

As our main theme has been the connection between filter stability and systems theory, we begin with an example of a systems-theoretic nature. Let $(X_k)_{k\ge0}$ be any stationary Markov chain, and consider additive noise observations

$$Y_k^\varepsilon = h(X_k) + \varepsilon\eta_k, \qquad (\eta_k)_{k\ge0} \text{ are i.i.d. } N(0,1). \quad (7)$$

Here h is an observation function and ε denotes the strength of the observation noise. Our basic question is: how well can we track the hidden process on the long run when the observation noise is small? To this end, consider the quantity

$$e(f) = \limsup_{\varepsilon\to0} \limsup_{k\to\infty} \mathbf{E}\big[\{f(X_k) - \mathbf{E}[f(X_k)|Y_1^\varepsilon,\ldots,Y_k^\varepsilon]\}^2\big].$$

We say that the filter achieves the *maximal accuracy property* if $e(f) = 0$ for any (square-integrable) function f. Of course, if the observation function h is invertible, it is clear that the maximal accuracy property should hold: indeed, in this case, the hidden process is revealed entirely as $\varepsilon \to 0$. On the other hand, when h is not one-to-one, it is far from obvious whether the maximal accuracy property can be achieved.

The practical relevance of this property to an engineer is roughly as follows. Suppose we are trying to track some hidden process, and we are trying to improve the precision of our tracking device. If the maximal accuracy property holds, it makes sense to invest effort into reducing the observation noise in the system. On the other hand, if the maximal accuracy property does not hold, there is only so much that can be achieved by reducing the noise: if we want to improve precision past a certain limit, we have no choice but to improve our detection hardware or add additional detectors in order to obtain more informative measurements.

For linear-Gaussian models, the maximal accuracy property was characterized in the work of Kwakernaak and Sivan [3]. However, their proofs once again rely heavily on the analysis of Kalman filtering equations (or rather, their dual control counterparts) and shed little light on the maximal accuracy problem for more general hidden Markov models. On the other hand, [27] gives a very general measure-theoretic characterization of the maximal accuracy property.

Perhaps most interesting is the special case where the hidden process X takes values in a finite state space, which we presently describe. Let $(X_k)_{k\ge0}$ be a stationary, finite state Markov chain with invariant measure λ, and let the observations $(Y_k^\varepsilon)_{k\ge1}$ be as in (7). Without loss of generality, we assume that λ has positive mass at each point. By a standard argument, we can extend the stationary measure $\mathbf{P}^\lambda$ in a canonical fashion such that the hidden process is defined

also for negative times $(X_k)_{k\in\mathbb{Z}}$. Moreover, as $\mu \ll \lambda$ for any μ by assumption, we can extend $\mathbf{P}^\mu$ to the the two-sided process $(X_k)_{k\in\mathbb{Z}}$ also by setting $d\mathbf{P}^\mu = \frac{d\mu}{d\lambda}(X_0)\, d\mathbf{P}^\lambda$.

Definition VI.1. A finite state hidden Markov model with additive noise observations (7) is called *reconstructible* if $\mathbf{P}^\mu|_{(h(X_k))_{k\le 0}} = \mathbf{P}^\nu|_{(h(X_k))_{k\le 0}}$ implies $\mu = \nu$.

Definition VI.2. A finite state hidden Markov model with additive noise observations (7) is said to satisfy the *graph coloring condition* if the following hold:

1) If $i \neq j$ and $\mathbf{P}[X_{k+1} = j | X_k = i] > 0$, then $h(i) \neq h(j)$.
2) If $i \neq j \neq k$ and both $\mathbf{P}[X_{k+1} = j | X_k = i] > 0$ and $\mathbf{P}[X_{k+1} = k | X_k = i] > 0$, then $h(j) \neq h(k)$.

Evidently reconstructibility is a sort of time-reversed observability property. The graph coloring condition is easily visualised if we draw the transition graph of the hidden process X and color each vertex according to the value of h at that point. We now obtain the following result.

Theorem VI.3. *The finite state hidden Markov model with additive noise observations (7) achieves the maximal accuracy property if and only if the model is reconstructible and satisfies the graph coloring condition.*

The proof of this odd result is given in [27] (in continuous time, but the proof is easily adapted to the discrete time setting). As one might imagine, the observability Theorem IV.2 plays a key role in establishing the result. One surprising observation is that, unlike in the case of filter stability, the necessary and sufficient condition for the maximal accuracy property in the finite state case is quite different in nature than the necessary and sufficient condition for the linear-Gaussian case. We refer to [27] for further discussion.

B. Uniform approximation of nonlinear filters

In the introduction, we explained the importance of the filter stability property as an issue of robustness: as the initial measure can typically not be estimated, there is no point in computing the filter unless it is robust to misspecification of the initial measure. However, in practice, other elements of the model, such as the transition probabilities of the hidden process or the observation structure, must also be calibrated to observed data. Though good statistical procedures exist to estimate these quantities, ultimately some small amount of model misspecification is inevitable. Moreover, in practice, we cannot compute the filter exactly and must therefore make numerical approximations. The effect of these errors on the performance of the filter in the long run is far from clear.

That an unfortunate choice of filter approximation can have disasterous consequences is illustrated in Figure 2. For simplicity, we have chosen a linear-Gaussian example where the filter can be computed exactly using the Kalman filtering equations. We now compare the exact filter with two different types of Monte Carlo approximations of the nonlinear filter using N particles.[5] Evidently one of the algorithms performs

[5]SIS stands for Sequential Importance Sampling, R stands for Resampling. The precise details of these algorithms are irrelevant to the present discussion. See [28], for example, and the references therein.

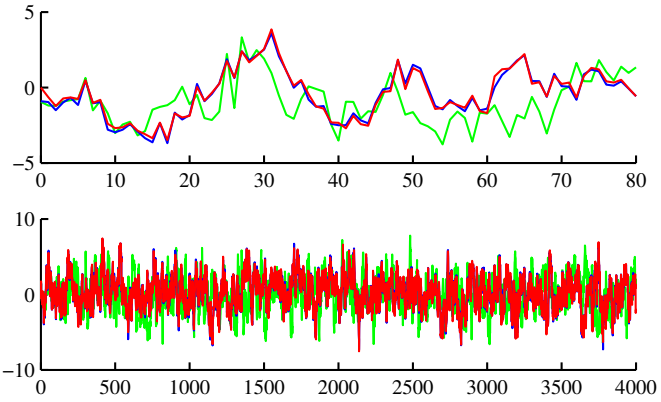

Fig. 2. SIS (green plot) and SIS-R (blue plot) Monte Carlo approximations, each using $N = 50$ particles, of the Kalman filter conditional mean (red plot) for the model $X_{k+1} = 0.9X_k + \xi_k$, $Y_k = X_k + \eta_k$. Both SIS and SIS-R converge to the exact filter as $N \to \infty$, but only SIS-R converges uniformly in time.

extremely poorly, while the other performs extremely well. Nonetheless, it is easy to prove that both algorithms converge to the exact filter as $N \to \infty$! Thus a naive convergence analysis gives us very little insight into the performance of filter approximations. However, Figure 2 suggests that we are asking the wrong question: note that the "bad" algorithm initially has small error, but the error accumulates over time, while the "good" algorithm has an approximation error which is independent of how long the algorithm has been running. The interesting question is therefore not whether an approximation simply converges, but whether it converges *uniformly* in time. This is a much more delicate issue.

Uniform approximation of nonlinear filters is known to be closely related to the filter stability property. This has been investigated, e.g., in [29], where even quantitative results are given. However, almost all results in the literature have been restricted to highly unrealistic assumptions. For example, the results of [29] rely on a uniform contraction property of the Markov kernel of the hidden process, which is almost never satisfied in applications. In contrast, using Theorem III.3 we can give a very general, albeit inherently qualitative, uniform approximation result for nonlinear filters.

To this end, we introduce a sequence $(\pi_k^N)_{k\ge 0}$ of filter approximations which are to converge to the exact filter $(\pi_k)_{k\ge 0}$ as $N \to \infty$ (for example, N could denote the number of particles used for Monte Carlo approximation, or $1/N$ could denote the degree of model misspecification). We further assume that the approximate filters are of *recursive type*: that is, $\pi_{k+1}^N = F^N[\pi_k^N, Y_{k+1}, \omega_{k+1}]$, where F^N is a suitably defined functional and $(\omega_k)_{k\ge 1}$ is an i.i.d. sequence that is independent of $(Y_k)_{k\ge 0}$ (the latter is needed to provide the additional randomness in Monte Carlo based approximations). It is sensible to consider approximations of recursive type, as it is well known that the filter itself can be computed in a recursive fashion: $\pi_{k+1} = F[\pi_k, Y_{k+1}]$, where F is defined by the usual prediction/Bayes update formula. The recursive structure guarantees that both $(X_k, \pi_k)_{k\ge 0}$ and $(X_k, \pi_k^N)_{k\ge 0}$ are (measure-valued) Markov chains with transition kernels Π and Π^N, respectively. Our general result now states that

if $\Pi^N \to \Pi$ in a suitable sense, and under the conditions of Theorem III.3, filter approximations of recursive type converge to the exact filter uniformly in a time average sense.

Theorem VI.4. *Suppose that the following hold:*

1) *Assumptions III.1 and III.2 are in force.*
2) $\Pi^N \to \Pi$ *as* $N \to \infty$ *uniformly on compacts.*
3) *The family* $\{\mathbf{E}[\pi_k^N] : k, N \geq 1\}$ *is tight.*

Then the filter approximation satisfies

$$\lim_{N\to\infty} \sup_{T\geq 1} \mathbf{E}\left[\frac{1}{T}\sum_{k=1}^{T} \|\pi_k^N - \pi_k\|_{\mathrm{BL}}\right] = 0,$$

i.e., the approximation converges uniformly in time average.

The proof of this result, which is based on a technique developed by Budhiraja and Kushner [30], is given in [31]. We refer to [31] for a more precise statement and discussion of the requisite assumptions. Let us note, however, that the filter stability result of Theorem III.3 is a key ingredient that allows us to obtain a result at this level of generality.

The main message of Theorem VI.4 is that under the assumptions needed for the filter stability Theorem III.3, almost any filter approximation of recursive type which converges in one time step to the exact filter will also converge uniformly in time. The setting is sufficiently general that one can consider both model misspecification and Monte Carlo approximations (or other numerical algorithms) within the same framework. In practice, the most difficult assumption to check is the (essentially technical) tightness assumption; it means, roughly speaking, that we must ensure that no mass can be "lost to infinity". For the case of SIS-R Monte Carlo approximations, several sufficient conditions for tightness are given in [31, section 4]. The SIS algorithm (cf. Figure 2), on the other hand, is not of recursive type in the sense required here: evidently the recursive nature is really essential in order to obtain useful filter approximations.

Finally, let us note that Theorem VI.4 holds only in the ergodic setting (Assumption III.1). Empirically, it seems that the filter can be approximated uniformly in time even in the absence of ergodicity if the model is sufficiently observable. No theoretical results to date support this observation, however, and the behavior of filter approximations in the nonergodic case therefore still remains a mystery.

Acknowledgments. It is a pleasure to acknowledge numerous conversations over the years with Pavel Chigansky, which have greatly stimulated my thinking about the problems discussed in this paper. I am grateful to Ofer Zeitouni for encouraging me to work on the maximal accuracy problem and for several very helpful discussions on this topic.

References

[1] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME Ser. D. J. Basic Engrg.*, vol. 82, pp. 35–45, 1960.

[2] R. S. Bucy and P. D. Joseph, *Filtering for stochastic processes with applications to guidance*. New York: Wiley-Interscience, 1968.

[3] H. Kwakernaak and R. Sivan, "The maximally achievable accuracy of linear optimal regulators and linear optimal filters," *IEEE Trans. Automatic Control*, vol. AC-17, pp. 79–86, 1972.

[4] D. Crisan and B. Rozovsky, Eds., *The Oxford University Handbook of Nonlinear Filtering*. Oxford University Press, 2010, to appear.

[5] D. Ocone and E. Pardoux, "Asymptotic stability of the optimal filter with respect to its initial condition," *SIAM J. Control Optim.*, vol. 34, pp. 226–243, 1996.

[6] R. van Handel, "Observability and nonlinear filtering," *Probab. Th. Rel. Fields*, vol. 145, pp. 35–74, 2009.

[7] P. Baxendale, P. Chigansky, and R. Liptser, "Asymptotic stability of the Wonham filter: Ergodic and nonergodic signals," *SIAM J. Control Optim.*, vol. 43, pp. 643–669, 2004.

[8] R. van Handel, "The stability of conditional Markov processes and Markov chains in random environments," *Ann. Probab.*, vol. 37, pp. 1876–1925, 2009.

[9] D. Revuz, *Markov chains*, 2nd ed. Amsterdam: North-Holland Publishing Co., 1984.

[10] A. Budhiraja, "Asymptotic stability, ergodicity and other asymptotic properties of the nonlinear filter," *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 39, pp. 919–941, 2003.

[11] P. Chigansky and R. van Handel, "A complete solution to Blackwell's unique ergodicity problem for hidden Markov chains," *Ann. Appl. Probab.*, 2010, to appear.

[12] H. von Weizsäcker, "Exchanging the order of taking suprema and countable intersections of σ-algebras," *Ann. Inst. H. Poincaré Sect. B (N.S.)*, vol. 19, pp. 91–100, 1983.

[13] L. Chaumont and M. Yor, *Exercises in probability*. Cambridge: Cambridge University Press, 2003.

[14] Y. G. Sinaĭ, "Kolmogorov's work on ergodic theory," *Ann. Probab.*, vol. 17, pp. 833–839, 1989.

[15] P. Masani, "Wiener's contributions to generalized harmonic analysis, prediction theory and filter theory," *Bull. Amer. Math. Soc.*, vol. 72, pp. 73–125, 1966.

[16] H. Kunita, "Asymptotic behavior of the nonlinear filtering errors of Markov processes," *J. Multivar. Anal.*, vol. 1, pp. 365–393, 1971.

[17] R. Cogburn, "On direct convergence and periodicity for transition probabilities of Markov chains in random environments," *Ann. Probab.*, vol. 18, pp. 642–654, 1990.

[18] R. van Handel, "Uniform observability of hidden Markov models and filter stability for unstable signals," *Ann. Appl. Probab.*, vol. 19, pp. 1172–1199, 2009.

[19] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. London: Springer-Verlag, 1993.

[20] R. van Handel, "Discrete time nonlinear filters with informative observations are stable," *Electr. Commun. Probab.*, vol. 13, pp. 562–575, 2008.

[21] D. Blackwell and L. Dubins, "Merging of opinions with increasing information," *Ann. Math. Statist.*, vol. 33, pp. 882–886, 1962.

[22] P. Chigansky and R. Liptser, "On a role of predictor in the filtering stability," *Electr. Commun. Probab.*, vol. 11, pp. 129–140, 2006.

[23] H. Kunita, "Supports of diffusion processes and controllability problems," in *Proceedings of the International Symposium on Stochastic Differential Equations (Res. Inst. Math. Sci., Kyoto Univ., Kyoto, 1976)*. New York: Wiley, 1978, pp. 163–185.

[24] J. R. Norris, *Markov chains*. Cambridge: Cambridge University Press, 1998.

[25] D. Blackwell and D. Freedman, "The tail σ-field of a Markov chain and a theorem of Orey," *Ann. Math. Statist.*, vol. 35, pp. 1291–1295, 1964.

[26] R. Isaac, "The tail σ-fields of recurrent Markov processes," *Apl. Mat.*, vol. 22, pp. 397–408, 1977.

[27] R. van Handel, "When do nonlinear filters achieve maximal accuracy?" *SIAM J. Control Optim.*, vol. 48, pp. 3151–3168, 2009.

[28] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*. New York: Springer, 2005.

[29] P. Del Moral, *Feynman-Kac formulae*. New York: Springer-Verlag, 2004.

[30] A. Budhiraja and H. J. Kushner, "Monte Carlo algorithms and asymptotic problems in nonlinear filtering," in *Stochastics in finite and infinite dimensions*. Boston: Birkhäuser, 2001, pp. 59–87.

[31] R. van Handel, "Uniform time average consistency of Monte Carlo particle filters," *Stoch. Proc. Appl.*, vol. 119, pp. 3835–3861, 2009.