Stochastic Observability and Filter Stability Under Several Criteria

Curtis McDonald and Serdar Yüksel, *Member, IEEE*

***Abstract*—Despite being a foundational concept of modern systems theory, there have been few studies on observability of nonlinear stochastic systems under partial observations. In this article, we introduce a definition of observability for stochastic nonlinear dynamical systems, which involves an explicit functional characterization. To justify its operational use, we establish that this definition implies filter stability under mild continuity conditions: an incorrectly initialized nonlinear filter is said to be stable if the filter eventually corrects itself with the arrival of new measurement information. Numerous examples are presented and a detailed comparison with the literature is reported. We also establish implications for various criteria for filter stability under several notions of convergence such as weak convergence, total variation, and relative entropy. These findings are connected to robustness and approximations in partially observed stochastic control.**

***Index Terms*—Filter stability, merging, nonlinear filtering, observability.**

I. Introduction

OBSERVABILITY is one of the most important and foundational concepts of modern systems and control theory with implications at the heart of its theory and applications [10], [30], [38], [39]. For deterministic linear systems, observability is defined as the exact recovery of any initial condition with measurements available until some finite time, and is characterized by an observability rank condition in both continuous time and discrete time [11]. For linear systems, such an observability definition is global (as it applies for all initial states) and is also directly applicable to stochastic counterparts of deterministic linear systems. For nonlinear systems, however, due to the challenges in the analysis, which prevent globality, the analysis is significantly more nuanced both for deterministic and stochastic setups. See Section II-D for a detailed discussion.

We study the stochastic setup in this article. Let us now introduce the probabilistic setup for a hidden Markov model (HMM) or partially observed Markov process (POMP). Let $(\mathcal{X},\mathcal{Y})$ be complete, separable, and metric (Polish) spaces equipped with their Borel sigma fields $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\mathcal{Y})$. $\mathcal{X}$ will be called the state space, and $\mathcal{Y}$ the measurement space. Let $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ be the set of probability measures on these spaces. Define the transition kernel T and measurement channel G as the mappings

$$\begin{aligned} T &: \mathcal{X} \to \mathcal{P}(\mathcal{X}) & G &: \mathcal{X} \to \mathcal{P}(\mathcal{Y}) \\ x &\mapsto T(dx'|x) & x &\mapsto G(dy|x). \end{aligned}$$

The system is initialized with a state $X_0 \in \mathcal{X}$ drawn from a prior measure μ on $(\mathcal{X},\mathcal{B}(\mathcal{X}))$. The state is then randomly updated via the transition kernel T, which makes the state process $\{X_n\}_{n=0}^{\infty}$ a Markov chain with initial measure μ and transition kernel T.

However, the state is not available at the observer, instead at time n the observer sees the observation Y_n where the conditional distribution of $Y_n|X_n$ is determined by the measurement channel G.

By stochastic realization arguments [23, Lemma 1.2], [9, Lemma 3.1], we can also view an equivalent construction of the system dynamics. Let $\{Z_n\}_{n=0}^{\infty}$ and $\{W_n\}_{n=0}^{\infty}$ be independent identically distributed (i.i.d.) $\mathcal{Z}$-valued noise processes, where $\mathcal{Z}$ can be taken to be [0,1] or $\mathbb{R}$ (or any other Polish space), without any loss of generality. Consider a partially observed dynamical system with the following model:

$$X_{n+1} = b(X_n, W_n) \tag{1}$$

$$Y_n = h(X_n, Z_n) \tag{2}$$

where W_n and Z_n can be assumed to take values from [0,1] or $\mathbb{R}$. Here, b defines the system dynamics and defines a transition kernel T for the Markov chain X_n. Assuming that Z_n has measure Q in $\mathcal{Z}$, the measurement function h defines the measurement channel G, which is the pushforward measure of Q under $h(x,\cdot)$. Throughout this article, we will work with either the general kernel and measurement channel notation T, G or with the specific functional form using b, h when convenient.

Thus, the observer needs to compute the conditional probability on the hidden variable X_n using the information available up to time $n \in \mathbb{Z}_+$. We have that $\{X_n, Y_n\}_{n=0}^{\infty}$ is a Markov chain, and we will denote P^{μ} as the probability measure on $\Omega = \mathcal{X}^{\mathbb{Z}_+} \times \mathcal{Y}^{\mathbb{Z}_+}$, endowed with the product topology, (and thus, $\omega \in \Omega$ is a sequence of states and measurements $\omega = \{(x_i, y_i)\}_{i=0}^{\infty}$), where $X_0 \sim \mu$. Such a stochastic system is referred to as a POMP (also called HMM) throughout this article.

Definition 1.1: We define the *one-step predictor* as the sequence

$$\pi_{n-}^{\mu}(\cdot) = P^{\mu}(X_n \in \cdot | Y_0, \ldots, Y_{n-1}), \quad n \in \mathbb{Z}_+$$

Manuscript received 27 April 2022; revised 2 May 2022, 27 November 2022, and 10 April 2023; accepted 17 July 2023. Date of publication 4 August 2023; date of current version 26 April 2024. This work was supported by the Natural Sciences and Engineering Research Council of Canada. An earlier version of this paper was presented at the 2018 Annual Allerton Conference [DOI: 10.1109/Allerton44049.2018]. Recommended by Associate Editor P. G. Mehta. *(Corresponding author: Serdar Yüksel.)*

Curtis McDonald is with the Department of Statistics and Data Science, Yale University, New Haven, CT 06520 USA (e-mail: curtis.mcdonald@yale.edu).

Serdar Yüksel is with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: yuksel@queensu.ca).

Digital Object Identifier 10.1109/TAC.2023.3302208

and we define the *nonlinear filter* as the sequence

$$\pi_n^\mu(\cdot) = P^\mu(X_n \in \cdot | Y_0, \ldots, Y_n), \quad n \in \mathbb{Z}_+.$$

Both of the above are regular conditional probability sequences defined on $\mathcal{X}$. We will use the notation $Y_{[0,n]} = Y_0, \ldots, Y_n$ to represent finite sets of random variables, and $Y_{[0,\infty)} = Y_0, Y_1, \ldots$ to represent infinite sequences. The recursive update equations for the filter or the predictor are known as the nonlinear filtering equations. Let us, for the time being, assume the existence of a likelihood function $g(x, y)$ for the measurement channel defined as follows. The measurement channel G is called dominated if there exists a reference measure λ such that $\forall x \in \mathcal{X}, G(Y_n \in \cdot | X_n = x) \ll \lambda$ where the notation "$\ll$" denotes absolutely continuity. We can then utilize a likelihood function $g(x, y) = \frac{dG(Y_n \in \cdot | X_n = x)}{d\lambda}(y)$ and write the filter π_{n+1}^μ recursively in terms of π_n^μ and $Y_{n+1} = y_{n+1}$ explicitly as a Bayesian update

$$\pi_{n+1}^\mu(dx_{n+1}) = F(\pi_n^\mu, y_{n+1})(dx_{n+1})$$

$$:= \frac{g(x_{n+1}, y_{n+1}) \int_{\mathcal{X}} T(dx_{n+1}|X_n = x)\pi_n^\mu(dx)}{\int_{\mathcal{X}} g(x_{n+1}, y_{n+1}) \int_{\mathcal{X}} T(dx_{n+1}|X_n = x)\pi_n^\mu(dx)}. \quad (3)$$

Suppose that an observer runs a nonlinear filter assuming that the initial prior is ν, when in reality the prior distribution is μ. The observer receives the measurements and computes the filter π_n^ν for each n, but the measurement process is generated according to the true measure μ.

The operational question for observability is that of *filter stability*, namely, if we have two different initial probability measures μ and ν, when do we have that the filter processes π_n^μ and π_n^ν merge in some appropriate sense as $n \to \infty$. In essence, when will our observations Y_n be informative enough to correct our incorrect prior ν and result in an accurate conditional measure for the hidden state.

The rest of this article is organized as follows. In Section I-A, notations and definitions are presented. In Section II, we present our main results. We present a detailed literature review after the statement of our main results in Section II-D. Examples of observable systems are given in Section III. Proofs are provided in Section IV. Finally, Section V concludes this article.

A. Notation and Preliminaries

Let $C_b(\mathcal{X})$ represent the set of continuous and bounded functions from $\mathcal{X} \to \mathbb{R}$.

Definition 1.2: Two sequences of probability measures P_n and Q_n merge weakly if $\forall f \in C_b(\mathcal{X})$ we have $\lim_{n\to\infty} |\int f dP_n - \int f dQ_n| = 0$.

Definition 1.3: For two probability measures P and Q, the total variation norm is defined as $\|P - Q\|_{TV} = \sup_{\|f\|_\infty \leq 1} |\int f dP - \int f dQ|$, where f is assumed measurable.

Note that merging in total variation implies weak merging since $C_b(\mathcal{X})$ is a subset of the set of measurable and bounded functions. We also utilize the relative entropy (Kullback–Leibler divergence) between two probability measures, although it is not a metric (since it is not symmetric).

Definition 1.4:

1) For two probability measures P and Q, we define the relative entropy as $D(P\|Q) = \int \log \frac{dP}{dQ} dP = \int \frac{dP}{dQ} \log \frac{dP}{dQ} dQ$, where $P \ll Q$ and $\frac{dP}{dQ}$ denotes the Radon–Nikodym derivative of P with respect to Q.
2) Let X and Y be two random variables, and let P and Q be two different joint measures for (X, Y) with $P \ll Q$. We define the (conditional) relative entropy between $P(X|Y)$ and $Q(X|Y)$ as

$$D(P(X|Y)\|Q(X|Y)) = \int \log\left(\frac{dP_{X|Y}}{dQ_{X|Y}}(x, y)\right) P(d(x, y))$$

$$= \int \left(\int \log \left(\frac{dP_{X|Y}}{dQ_{X|Y}}(x, y) \right) P(dx|Y = y) \right) P(dy). \quad (4)$$

Some notational discussion is in order. For some probability measures, such as $P^\mu(Y_{[0,n]} \in \cdot)$ or $P^\mu(X_n \in \cdot)$, it will be convenient to denote the random variable inside the measure and take out the set argument. When we take the relative entropy of such measures, to make the notation shorter, we will drop the "$\in \cdot$" argument and write $D(P^\mu(Y_{[0,n]})\|P^\nu(Y_{[0,n]}))$.

Note that in a conditional relative entropy, we are integrating the logarithm of the Radon–Nikodym derivative of the conditional measures $P(X|Y)$ and $Q(X|Y)$ over the joint measure of P on (X,Y). The second equality (4) shows that this can be thought of as the expectation of the relative entropy $D(P(X|Y = y)\|Q(X|Y = y))$ at specific realizations of $Y = y$, where the expectation is over the marginal measure of P on Y. When we apply this to the filter, π_n^μ and π_n^ν are realizations of the filter for specific measurements; therefore, when we discuss their relative entropy, we take the expectation over the marginal of P^μ on $Y_{[0,n]}$. We write this as $E^\mu[D(\pi_n^\mu\|\pi_n^\nu)]$, where $D(\pi_n^\mu\|\pi_n^\nu)$ plays the role of the inner integral in (4).

We now introduce some additional notation that will be useful when dealing with sigma fields rather than random variables directly. Strictly speaking, we have two probability measures P^μ and P^ν on $(\mathcal{X}^{\mathbb{Z}_+} \times \mathcal{Y}^{\mathbb{Z}_+}, \mathcal{B}(\mathcal{X}^{\mathbb{Z}_+} \times \mathcal{Y}^{\mathbb{Z}_+}))$. We denote by $\mathcal{F}_{a,b}^{\mathcal{X}}$ the sigma field generated by $(X_a, \ldots, X_b)$ and similarly for $\mathcal{Y}$. We also write $\mathcal{F}_n^{\mathcal{X}}$ for the sigma field generated by X_n. We then have $\mathcal{F}_{0,\infty}^{\mathcal{X}} \vee \mathcal{F}_{0,\infty}^{\mathcal{Y}}$ as the sigma field generated by all state and measurement sequences. When we write $P^\mu(X_{[0,n]})$ we are discussing the measure P^μ restricted to the sigma field $\mathcal{F}_{0,n}^{\mathcal{X}}$, which we will denote $P^\mu|_{\mathcal{F}_{0,n}^{\mathcal{X}}}$. Similarly, for some set $A \in \mathcal{F}_{0,\infty}^{\mathcal{X}} \vee \mathcal{F}_{0,\infty}^{\mathcal{Y}}$, we write $P^\mu((X_{[0,\infty)}, Y_{[0,\infty)}) \in A|Y_{[0,n]})$ as the conditional measure of P^μ with respect to the sigma field $\mathcal{F}_{0,n}^{\mathcal{Y}}$, which we denote $P^\mu|\mathcal{F}_{0,n}^{\mathcal{Y}}$. We can also consider restricting and conditioning simultaneously, this for example is the case with the nonlinear filter: $\pi_n^\mu(\cdot) = P^\mu(X_n \in \cdot|Y_{[0,n]}) = P^\mu|_{\mathcal{F}_n^{\mathcal{X}}}|\mathcal{F}_{0,n}^{\mathcal{Y}}$.

The key relationship between relative entropy and total variation is Pinsker's inequality (see, e.g., [16]), which states that for two probability measures P and Q, we have that $\|P - Q\|_{TV} \leq \sqrt{\frac{2}{\log(e)} D(P\|Q)}$.

Criteria for stability: We note the following definitions for filters, but they can also be defined for predictors.

Definition 1.5:

1) A filter process is stable in the sense of weak merging in expectation if for any $f \in C_b(\mathcal{X})$ and any prior ν with $\mu \ll \nu$, we have

$$\lim_{n\to\infty} E^\mu \left[\left| \int f d\pi_n^\mu - \int f d\pi_n^\nu \right| \right] = 0.$$

2) A filter process is stable in the sense of weak merging P^μ almost surely (a.s.) if there exists a set of measurement sequences $A \subset \mathcal{Y}^{\mathbb{Z}_+}$ with P^μ probability 1 such that for any sequence in A, for any $f \in C_b(\mathcal{X})$ and any prior ν with $\mu \ll \nu$, we have

$$\lim_{n\to\infty}\left|\int f d\pi_n^\mu - \int f d\pi_n^\nu\right| = 0.$$

3) A filter process is stable in the sense of total variation in expectation if for any measure ν with $\mu \ll \nu$, we have

$$\lim_{n\to\infty} E^\mu[\|\pi_n^\mu - \pi_n^\nu\|_{TV}] = 0.$$

4) A filter process is stable in the sense of total variation P^μ a.s. if for any measure ν with $\mu \ll \nu$ we have

$$\lim_{n\to\infty} \|\pi_n^\mu - \pi_n^\nu\|_{TV} = 0 \;\; P^\mu \text{ a.s.}$$

5) A filter process is stable in relative entropy if for any measure ν with $\mu \ll \nu$

$$\lim_{n\to\infty} E^\mu[D(\pi_n^\mu \| \pi_n^\nu)] = 0.$$

6) For $f : \mathcal{X} \to \mathbb{R}$, define the Lipschitz norm

$$\|f\|_L = \sup\left\{\frac{|f(x)-f(y)|}{d(x,y)}\,\middle|\, d(x,y)\neq 0\right\}.$$

With $\mathrm{BLip} := \{f : \|f\|_L \le 1, \|f\|_\infty \le 1\} \subset C_b(\mathcal{X})$, we define the bounded Lipschitz (BL) metric as

$$\|P-Q\|_{\mathrm{BL}} = \sup_{f\in\mathrm{BLip}}\left|\int f dP - \int f dQ\right|.$$

A system is then stable in the sense of BL-merging P^μ a.s. if we have $\|\pi_n^\mu - \pi_n^\nu\|_{BL} \to 0 \;\; P^\mu$ a.s.

We note that *merging* of probability measures is different from the *convergence* of a sequence of probability measures to a limit measure. In convergence, we have some sequence P_n and a static limit measure P; in merging, we have two sequences P_n and Q_n, which may not individually have limits, but come closer together for large n in one of the merging notions defined previously [17].

II. Statement of the Main Results and Literature Review

A. Stochastic Nonlinear Observability

We first introduce our notion of an observable system.

Definition 2.1:

1) *One-Step Observability:* A POMP is said to be one-step observable if for every $f \in C_b(\mathcal{X}), \epsilon > 0, \exists$ a measurable and bounded function $g : \mathcal{Y} \to \mathbb{R}$ such that

$$\left\| f(\cdot) - \int_Y g(y)G(dy|\cdot)\right\|_\infty < \epsilon.$$

2) *N-Step Observability:* A POMP is said to be N-step observable if for every $f \in C_b(\mathcal{X}), \epsilon > 0, \exists$ a measurable and bounded function $g : \mathcal{Y}^N \to \mathbb{R}$ such that

$$\left\| f(\cdot) - \int_{\mathcal{Y}^N} g(y_{[1,N]})P(dy_{[1,N]}|X_1 = \cdot)\right\|_\infty < \epsilon \quad (5)$$

where we note that the conditional probability $P(dy_{[1,N]}|X_1 = x_1)$ is independent of the prior measure.

3) *Observability:* A POMP is *observable* if for every $f \in C_b(\mathcal{X})$ and every $\epsilon > 0$, there exist $N \in \mathbb{N}$ and a measurable and bounded function g (both possibly dependent on f and ϵ) such that (5) applies. Note that if a POMP is N step observable for some finite $N \in \mathbb{N}$, then it is *observable*, but the reverse implication is not necessarily the case.

A number of remarks are in order.

Remark 2.2: In the definition above, we can instead of $C_b(\mathcal{X})$ consider any dense subset in $C_b(\mathcal{X})$. For example, if $\mathcal{X}$ is a compact subset of $\mathbb{R}^k$, we can consider polynomials as these form a dense subset, or we can consider smooth functions defined on $\mathcal{X}$, or functions that are expressed as linear combinations of harmonics, Haar wavelets, etc. An example is provided in Section III-B.

Remark 2.3 (One-step observability and universality in the controlled setup): The definition of one-step observability is a specific case of N-step observability; however, the distinction is useful for at least two reasons: 1) one-step observability is often easier to check since one does not need to consider the effect of the state transition kernel, as this definition is only concerned with the measurement channel itself. On the other hand, there exist many setups where a system is observable, but not one-step observable; see e.g. Section III-A. 2) Even though in this article, we consider a control-free setup, in a controlled context studied in a companion paper [44], it follows that one-step observability would be independent of any control policy (that is, observability would be universal over all policies and associated filter stability results apply under any control policy), but $N > 1$ step observability would be handled much more cautiously as this condition would be dependent on the control policy adopted. Recently, filter stability results have been shown to be consequential in showing near-optimality of finite memory control policies and associated learning theoretic results for partially observed Markov decision processes (POMDPs) [33], [34]. Accordingly, one-step observability results are particularly applicable for such scenarios.

Remark 2.4: If the measurement kernel satisfies an absolute continuity condition so that $G(dy|x) = h(x,y)\lambda(dy)$ and if there exists a finite measure K such that $G(dy|x) \le K(dy)$ (so that the family of kernels $\{G(dy|x), x \in \mathcal{X}\}$ is majorized by K leading to a uniformly countable additive family of measures), then by Lusin's theorem [18, Th. 7.5.2] and the extension theorem of Tietze [19, Th. 4.1], we can replace g in the above with a continuous function g_c. The relaxation to such continuous g_c is useful when one would like to approximate the channels with those that are quantized. This then leads to an easier way to test observability via a rank condition, e.g., when $\mathcal{X}$ is finite; see Section III-A.

Remark 2.5: One should note that the definition is not one of invertibility; it only requires that there exists some g and N such that the error between the conditional expectation of $g(y_{[1,N]})$, given $X_1 = x$, and $f(x)$ is small. In particular, X_1 is not necessarily, even approximately, recoverable given the measurements. Invertibility, however, would be a special case being a sufficient condition, as we will see in the examples.

Remark 2.6 (Recovery of Initial Probability Measure): By our definition of observability, for every $f \in C_b(\mathbb{X})$, the value is $\langle \mu, f\rangle := E_\mu[f(X)]$ determinable with arbitrary precision by the measurements (since f is recovered, uniformly over a given compact set, with arbitrary precision). Since a countable collection of continuous and bounded functions uniquely distinguish probability measures [20, Th. 3.4.5] (that is, such continuous and bounded functions form a *separating class*, see also [6, p. 13]), this amounts to the recovery of the initial probability measure as more and more measurements are collected. This then leads to the conclusion that our definition implies van Handel's definition given in (7), noted further below (note that this also applies for noncompact setups under Definition 2.14 as every individual probability measure is tight).

B. Filter Stability Under the Observability Definition

The presented observability definition leads to predictor stability in the following sense.

Theorem 2.7: Let

$$P^{\mu}|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}} \ll P^{\nu}|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}. \tag{6}$$

If the POMP is observable, then π^{μ}_{n-} and π^{ν}_{n-} merge weakly as $n \to \infty$, P^{μ} a.s.

A sufficient condition for (6) is that the priors satisfy $\mu \ll \nu$. The following assumption will allow us to use the recent results in [31] (see also [21]) and conclude the weak merging of the filter in expectation from the almost sure convergence of the predictor.

Assumption 2.8: The measurement channel is continuous in total variation. That is, for any sequence a_n with $\lim_{n\to\infty} a_n = a \in \mathcal{X}$, we have $\|G(\cdot|X_0 = a_n) - G(\cdot|X_0 = a)\|_{TV} \to 0$.

An example of a channel, which is continuous in total variation, is as follows [31, Sec. 2.2]: $Y_n = F(X_n) + W_n$, where F is continuous and W_n admits a continuous density function (such as a Gaussian), where an analysis based on convolution and Scheffé's lemma leads to the conclusion.

Theorem 2.9: Let Assumption 2.8 hold, if the predictor merges weakly P^{μ} a.s., then the filter merges weakly in expectation.

1) Localized Observability for Noncompact Signal Spaces: While the definition of observability that we introduced is valid for both compact and noncompact state spaces, it may be difficult to satisfy the definition in a noncompact state space with a uniform bound on the approximation error. This will be relaxed in the following, where we assume $\mathcal{X}$ to be Euclidean.

Definition 2.10: Given a compact set K, a POMP is called K locally predictable if there exists a sequence of $\mathcal{F}^{\mathcal{Y}}_{0,n-1}$ (with $n \in \mathbb{N}$) measurable mappings (random variables) $a_n : \mathcal{Y}^n \to \mathcal{X}$ such that

$$\pi^{\nu}_{n-}(K + a_n) = 1 \; P^{\mu} \text{ a.s.}$$

for every $\mu \ll \nu$.

This definition can be interpreted as follows. Regardless of the prior ν, upon seeing observations $Y_{[0,n-1]}$, we can be sure X_n lives in a compact set $K_n = K + a_n$. We can think of a_n as a "centering" value based on the observations $Y_{[0,n-1]}$ and K as the compact set around this centering value, in which X_n must live conditioned on observations $Y_{[0,n-1]}$. This is then paired with a definition of local observability.

Definition 2.11: Given a compact set K, a POMP is called K locally observable if for every continuous and bounded function f, every sequence of numbers a_n, and every $\epsilon > 0$, there exists a sequence of uniformly bounded measurable functions g_n such that

$$\sup_{x \in K + a_n} \left| f(x) - \int_{\mathcal{Y}} g_n(y) G(dy|x) \right| \le \epsilon$$

for every $n \in \mathbb{N}$.

Theorem 2.12: Assume $\mu \ll \nu$ and that there exists a compact set K such that the POMP is K locally predictable and K locally observable. Then, the predictor merges weakly P^{μ} a.s.

The result above is intuitive as we can specify the compact set K, and the shifted sets $K_n = K + a_n$, over which we must approximate the function. However, the definitions can also be constructed taking a more relaxed approach and appealing to tightness rather than a probability one statement, but in this case it is more difficult to satisfy the local definition of observability.

Definition 2.13: A POMP is called *locally predictable* if there exists a sequence of $\mathcal{F}^{\mathcal{Y}}_{0,n-1}$ (with $n \in \mathbb{N}$) measurable mappings $a_n : \mathcal{Y}^n \to \mathcal{X}$ such that the family of measures

$$\tilde{\pi}^{\nu}_{n-}(\cdot) := \pi^{\nu}_{n-}(\cdot + a_n)$$

for every $\mu \ll \nu$, is a uniformly tight family of measures.

Definition 2.14: A POMP is called locally observable if for every continuous and bounded function f, every compact set K, every sequence of numbers a_n, and every $\epsilon > 0$, there exists a sequence of uniformly bounded measurable functions g_n such that

$$\sup_{x \in K + a_n} \left| f(x) - \int_{\mathcal{Y}} g_n(y) G(dy|x) \right| \le \epsilon$$

$$\sup_{x \notin K + a_n} \left| \int_{\mathcal{Y}} g_n(y) G(dy|x) \right| \le 2\|f\|_{\infty}$$

for every $n \in \mathbb{N}$.

Theorem 2.15: Assume $\mu \ll \nu$ and that the POMP is locally predictable and locally observable. Then, the predictor merges weakly P^{μ} a.s.

An example is given in Section III-D.

C. Relations Between Various Criteria for Filter Stability

It follows from Pinsker's inequality that relative entropy merging implies total variation merging, which in turn implies weak merging (by Definitions 1.2 and 1.3). In this section, we are interested in conditions for when the converse direction holds, i.e., weak merging implies total variation or relative entropy merging. Recall the definition that for the measurement channel $G(Y_n \in \cdot|X_n = x)$ to be *dominated* in the sense that there exists a reference measure λ such that $\forall x \in \mathcal{X}, G(dy|x_n = x) \ll \lambda$. Then, we define the Radon–Nikodym derivative

$$g(x, y) := \frac{dG(y_n \in \cdot|x_n = x)}{d\lambda(\cdot)}(y)$$

which serves as a likelihood function (a conditional probability density function).

Assumption 2.16:

1) $T(\cdot|x)$ is absolutely continuous with respect to a dominating measure ϕ for every $x \in \mathcal{X}$, so that $t(x_1, x) = \frac{dT(\cdot|x)}{d\phi}(x_1)$ where t is continuous in x for every $x_1 \in \mathcal{X}$.
2) $g(x, y)$ is bounded and continuous in x for every fixed y. Furthermore, $g(x, y) > 0$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$.

Assumption 2.17: $T(\cdot|x)$ is absolutely continuous with respect to a dominating measure ϕ for every $x \in \mathcal{X}$, so that

$t(x_1, x) = \frac{dT(\cdot|x)}{d\phi}(x_1)$. The family of (conditional densities) $\{t(\cdot, x)\}_{x\in\mathcal{X}}$ is uniformly bounded and equicontinuous.

Theorem 2.18: Let $\mu \ll \nu$. Let either one of the following hold.

i) Assumption 2.16.

ii) Assumption 2.17.

Then, if the predictor is stable in the weak sense P^μ a.s., then it is also stable in total variation P^μ a.s.

Since the total variation of any two probability measures is uniformly bounded, stability in the almost sure sense implies that in expectation (with the same also holding for predictors). Thus, Theorem 2.18 also presents condition for predictor merging in total variation in expectation.

Theorem 2.19: The filter merges in total variation in expectation if and only if the predictor merges in total variation in expectation.

Recently, filter stability results under total variation in expectation (as in [43]) have been shown to be consequential in showing the optimality of finite memory control policies in POMDPs (see [34, Sec. 4.3 and Th. 9] and [33, Th. 3.2, 3.3, and 4.1]).

Theorem 2.20: Assume there exists some finite n such that $E^\mu[D(\pi_n^\mu \| \pi_n^\nu)] < \infty$ and some m such that $D(P^\mu|_{\mathcal{F}^y_{0,m}} \| P^\nu|_{\mathcal{F}^y_{0,m}}) < \infty$. Then, the filter is stable in relative entropy if and only if it is stable in total variation in expectation.

We note that both of the conditions on the finiteness of relative entropies in Theorem 2.20 are minor and hold for example if $D(\mu\|\nu) < \infty$. In the special setup where the measurement sigma field is the trivial one (with no information), or more generally Y_n is independent of X_n; the above recovers the following result, which generalizes Barron [3], [4] and Fritz [22], who had established the relative entropy convergence of sequences of probability measures for each time stage to the invariant measure for reversible Markov chains. This result also generalizes Theorem 5 of Harremoës and Holst [26], which considers countable state space chains with a uniform irreducibility assumption.

Theorem 2.21: Let X_t be a Markov chain with π being its unique invariant probability measure. Let π_t denote the measure $P^{\pi_0}(X_t \in \cdot)$, where $X_0 \sim \pi_0$. Let $\pi_t \to \pi$ in total variation. If $D(\pi_{t_0}||\pi) < \infty$ for some $t_0 < \infty$, then

$$D(\pi_t||\pi) \downarrow 0.$$

In particular, for an aperiodic positive Harris recurrent Markov chain, if $D(\pi_{t_0}||\pi) < \infty$ for some $t_0 < \infty$, then $D(\pi_t||\pi) \downarrow 0$.

Proof: The proof follows directly from Theorem 2.20. In the special case of positive Harris recurrence, the result follows since for aperiodic positive Harris recurrent Markov chains, $\pi_t \to \pi$ in total variation (see [45, Th. 13.0.1]).

D. Literature Review and Comparison of Results

For deterministic linear systems, exact recovery of any initial condition with measurements available until some finite time is defined as observability and is characterized by an observability rank condition in both continuous time and discrete time [11]. For linear systems, such an observability definition is global (as it applies to all initial states) and is universal in the control policies applied, as the control policy does not affect the estimation errors (known as the *no-dual effect* [2] property). For nonlinear systems, however, due to the challenges in the analysis, which prevent globality as well as controldependence, more modest and localized definitions are to be imposed: for deterministic continuous-time nonlinear systems [28] and [50] present local indistinguishability conditions with subtle differences, and establish relations with Lie-theoretic characterizations, which generalize observability rank conditions for nonlinear systems defined locally. For discrete-time deterministic models, observability has also been defined by invertibility or exact recovery of an initial state, locally, given measurements with finitely many observations. Nijmeier [46] developed discrete-time analogues of the observability notions presented in [28] (see also [50] for sampled continuous-time systems). The authors in [41] and [42] introduced a nonlinear stochastic observability definition through entropy, where the conditional entropy of the hidden state given measurements not being the same as the unconditional entropy implies observability. Ugrinoovski [52] also presented an information theoretic formulation, and defines observability as an informativeness condition.

In the filtering literature for control systems, the classical setup involves the linear Gaussian system. The filter in this case is the celebrated Kalman filter, where the finite-dimensional Kalman filter is computed recursively using the Riccati equation. Under linear observability and controllability conditions, the Riccati equation admits a unique solution [10], [38], [39], which is the unique limit of the Riccati recursions regardless of the initialization. Thus, the Kalman filter is stable with respect to incorrect, though still Gaussian, priors under the aforementioned conditions (we note that partial convergence and robustness results on the asymptotic equivalence of conditional expectations and linear estimates for non-Gaussian priors for linear systems are reported in [51]). The time-varying linear system setup has been studied in [1].

In a recent paper, we studied the implications on filter stability in robust control [44]. Implications of filter stability (such as the results reported in [43]) on finite memory approximations in optimal stochastic control have been presented in [33] and [34]. It is worth pointing out that there has been a recurrent theme on the duality between controllability and observability; for a recent work in this direction, see [35] and [36]. Filter stability for deterministic systems under noisy measurements has recently been studied in [48].

A strict version of our observability definition is captured in [13, eq. 1.7]. The idea there is to express, exactly, a continuous function $f(x)$ by integrating a measurable function $g(y)$ over the conditional distribution for Y given $X = x$. A fundamental result that pairs with observability is that of Blackwell and Dubins [7], an implication of which [13] independently arrived at. Blackwell and Dubins [7] used martingale convergence theorem to show that if P and Q are two measures on a fully observed stochastic process $\{X_n\}_{n=0}^\infty$ with $P \ll Q$, then the conditional distributions on the future based on the past merge in total variation P a.s., that is, P a.s.

$$\|P(X_{[n+1,\infty)} \in \cdot|X_{[0,n]}) - Q(X_{[n+1,\infty)} \in \cdot|X_{[0,n]})\|_{TV} \to 0.$$

van Handel [54] introduced a definition of observability for POMPs. Namely, a system is observable if every prior results in a unique probability measure on the measurement sequences

$$P^\mu|_{\mathcal{F}^y_{0,\infty}} = P^\nu|_{\mathcal{F}^y_{0,\infty}} \implies \mu = \nu. \tag{7}$$

van Handel [54] showed that the above leads to filter stability for continuous-time models with compact state space. van

Handel [56] extended these results to noncompact state spaces, where *uniform observability* is introduced. The result of Blackwell and Dubins [7] is utilized to show that uniform observability would imply filter stability in BL distance [53]. Nonetheless, this condition is implicit; van Handel [53] only studied the measurement channel where $h(x,z) = f(x) + z$, where f^{-1} is uniformly continuous and Z must have an everywhere nonzero characteristic function (e.g., a Gaussian distribution). For a compact state space, van Handel [56] established that uniform observability and observability are equivalent notions. We also note that for a finite state space with a nondegenerate measurement channel (i.e., likelihood function $g(x,y) > 0$), stability can be fully characterized via observability and a detectability condition [54], [57, Th. V.2], or [12, Ths. 2.7 and 3.1].

As noted in Remark 2.6, our definition implies (7), which is a statement of invertibility with no clear guidance on how to test this property; our definition is explicitly given in a test function formulation, making it more interpretable and easier to apply to various systems of interest. In addition, in the work studied here, we consider discrete time processes, and thus, the predictor and the filter are distinct objects. Our definition of observability only implies the weak merging of the predictor a.s., not the filter directly. Conditions are needed to relate the merging of the predictor to that of the filter. This is also addressed in this article building also on recent results on the regularity properties of nonlinear filters from [31] (see also [21]).

In an early work by Kunita [37], the stability of the filter process is studied in light of the limit sigma fields of the processes (e.g., $\mathcal{F}^{\mathcal{Y}}_{0,\infty}$ and $\mathcal{F}^{\mathcal{X}}_{0,\infty}$). Kunita's work unfortunately made a technical error on the exchange of orders in supremum and intersection operations on sigma fields: a concise derivation of the corrected result is presented in [14, eq. 1.10]; here, we are presented with a sufficient and necessary condition for the merging of the filter in total variation in expectation based on comparing the sigma fields $\mathcal{F}^{\mathcal{Y}}_{0,\infty}$ and $\bigcap_{n\geq 0} \mathcal{F}^{\mathcal{X}}_{n,\infty} \vee \mathcal{F}^{\mathcal{Y}}_{0,\infty}$. That is, the filter merges in total variation in expectation if and only if P^μ a.s. :

$$E^\nu\left[\frac{d\mu}{d\nu}(X_0)|\mathcal{F}^{\mathcal{Y}}_{0,\infty}\right] = E^\nu\left[\frac{d\mu}{d\nu}(X_0)|\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty} \vee \mathcal{F}^{\mathcal{Y}}_{0,\infty}\right].$$

Relative entropy as a measure of discrepancy between the true filter and the incorrectly initialized filter is studied by Clark et al. [15]. Here, the authors considered the filtering problem in continuous time and with a dominated measurement channel. The authors established the relative entropy of the true filter and the incorrect filter as a supermartingale, and its convergence to a limit. However, the paper did not establish the convergence to zero. A notable setup where actual convergence (of the relative entropy) to zero is established is the (rather specific) Beneš filter studied in [47]. This problem also has relations to the relative entropy convergence of Markov chains to invariance: in the case where the measurements are trivial, the convergence problem reduces to what has been studied in [3], [4], [22], [26], and[27] on relative entropy convergence of Markov chains to invariant measures.

Contributions and comparison with the literature: In view of the review above, our contributions are as follows.

1) *Stochastic observability:* In Section II-A, we present a definition of stochastic observability. This definition is functionally explicit and testable, and due to its functional approximation characterization, it allows various analytical methods to be applicable for verification (see Remarks 2.2–2.5).

   Under this definition, we establish predictor stability (in the weak convergence/merging sense). We note that observability, for the discrete time case as studied here, only implies weak merging of the predictor a.s., not the filter directly. This is addressed in this article building also on recent results on the regularity properties of nonlinear filters from [31] (see also [21]). We also note that the Blackwell and Dubins theory of merging on which our approach builds (similar to [54] and [56]) applies to infinite sequences of future events (i.e., $P^\mu(Y_{[n+1,\infty)}|Y_{[0,n]})$), this is utilized in our definition of N step observability leading to application examples of broad generality. Accordingly, our definition is not only a function of the measurement channel, but also of the system dynamics; unlike some related results in the literature.

   In addition, we provide several examples in Section III.

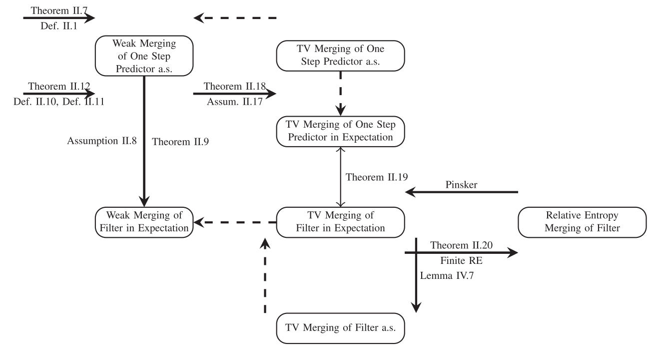

Fig. 1. Flow of ideas and conditions for filter stability.

2) *On various convergence and merging criteria:* We establish new results relating various criteria for filter stability (as depicted in Fig. 1), independent of the mechanism used to arrive at filter stability: we study filter stability under weak merging and total variation merging in expectation and a.s., as well as relative entropy. In Section II-C,

   a) We place mild assumptions on the transition/measurement kernels to extend weak merging of the predictor to total variation merging.

   b) We show that total variation merging of the predictor and filter are equivalent.

   c) Under a mild finiteness condition on the relative entropy sequence, we also establish equivalence of relative entropy merging and total variation merging.

Using the chain rule for relative entropy, the relative entropy error was shown to be a nonincreasing sequence by Clark et.al. [15], but *its convergence to zero was not established*, except for the specific case of the Beneš filter in [47]. Theorem 2.20 establishes the equivalence between relative entropy merging and total variation, and thus, convergence of the relative entropy error to zero is proven here (we note that this is a result that is hinted at in the literature, see [14, Remark 4.2] or [55, Remark 5.9] but not explicitly proven). This result applies to setups beyond filter stability: in the case where the measurements are trivial, Theorem 2.21 generalizes [3], [4], [22], [26] on relative entropy convergence of Markov chains to invariant measures, where the first references due to Barron and Fritz had considered reversible Markov chains and the latter due to Harremoës and Holst focused on countable state Markov chains under a uniform irreducibility assumption. On Theorem 2.19, we note first that much of the literature focuses on continuous time, where the predictor is not used in the analysis. In discrete time, [53, Lemma 4.2] proves that the merging of the predictor in total variation in expectation implies that of the filter. However, this result relies on a domination assumption in the measurement channel and the specific structure of the filter recursion equation [14, eq. 1.4]. Theorem 2.19 is, accordingly, a more general result.

3) *Implications to near optimality of finite window policies in POMDPs:* Our findings lead to practically relevant and mathematically consequential implications to robustness and approximations for controlled partially observable models; i.e., POMDPs: McDonald and Yüksel [44] had studied controlled filter stability where it was shown that one-step observability introduced here leads to stochastic observability universally over all admissible control policies, which then leads to refined robustness results when compared with [32]. In this article, we consider the control-free case, which allows us to consider N-step observability, with $N > 1$. In addition, we present numerous explicit examples, which, in the one-step observable setup, are then directly applicable to such robustness results. Recently, filter stability results under total variation (as well as weak convergence under slightly more restrictive setups) have been shown to be consequential in showing the optimality of finite memory control policies in POMDPs; see [34, Sec. 4.3 and Th. 9] and [33, Ths. 3.2, 3.3, and 4.1], where connections with weak merging and total variation merging are made explicit in the approximation error bounds (see [29] for an earlier study where the dependence on filter stability is implicit; further related recent studies include [24]). Accordingly, the results in this article, notably Theorems 2.18 and 2.19, are directly applicable in showing that with merging under total variation in expectation, one can show that optimal policies for POMDPs can be approximated by those which use only finite window of measurements and control actions.

III. Observable System and Measurement Channel Examples

We note that in this section, it will be more convenient to describe our measurement channels via the equivalent functional realization [see (1)], with explicit noise variable Z_n and a measurement function $Y_n = h(X_n, Z_n)$, and thus, this will the convention we will use to define the measurement channel G for the examples presented in the following.

A. Finite State and Noise Space

Consider a finite setup $\mathcal{X} = \{a_1, \dots, a_n\}$ and $\mathcal{Z} = \{b_1, \dots, b_m\}$. Now, assume $h(x, z)$ has K distinct outputs, where $1 \leq K \leq (n)(m)$ and $\mathcal{Y} = \{c_1, \dots, c_K\}$. We note that for such a setup, there is already a sufficient and necessary condition provided in [57, Th. V.2]. However, we examine this case to show that our definition is equivalent to the sufficient direction of this theorem, which is the notion of observability presented in [54].

For each x, $h_x(\cdot) := h(x, \cdot)$ can be viewed as a partition of $\mathcal{Z}$, assigning each $b_i \in \mathcal{Z}$ to an output level $c_j \in \mathcal{Y}$. We can track this by the matrix $H_x(i, j) = 1$ if $h_x(b_i) = c_j$ and zero else. Let Q be the $1 \times m$ vector representing the probability measure of the noise. Let us first consider one-step observability. Let $g(c_i) = \alpha_i$, with $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{bmatrix}$, and $\int_{\mathcal{Z}} g(h(x, z)) Q(dz)$ $=: QH_x\alpha$. Therefore, any function $f(x)$ can be expressed as an $n \times 1$ vector, and hence, the question reduces to finding a vector α so that $f = QH\alpha$, and the system is one-step observable if and only if the matrix $A \equiv \begin{pmatrix} QH_{a_1} \\ \vdots \\ QH_{a_n} \end{pmatrix}$ is rank n. Consider then N step observability. We wish to solve equations of the form

$$f(x) = \int_{\mathcal{Y}^N} g(y_{[1,N]}) dP^\mu(y_{[1,N]} | x_1 = x). \tag{8}$$

With knowledge of $Q, h(\cdot, \cdot)$, and T we can directly compute the transition kernel for the joint measure $Y_{[1,n]} | X_1$; however, the size of this matrix is n by K^n, where $|\mathcal{X}| = n$ and $|\mathcal{Y}| = K$, so complexity grows exponentially. We can deduce a sufficient, but not necessary, condition for n-step observability using the marginal conditional measures. Consider that $P^\mu(y_k \in \cdot | X_1 = a_j) = T(a_j| :)T^{k-2}A$, $k \geq 2$ where $T(a_j| :)$ represents the jth row of the transition matrix. Note that these are all $1 \times K$ vectors and represent the marginal measures of $Y_k | X_1$. Consider the class of functions $\mathcal{G}^n = \{g : \mathcal{Y}^n \to \mathbb{R}\}$ and a subclass $\mathcal{G}^n_{\text{LC}} = \{g(y_{[1,n]}) = \sum_{i=1}^n \alpha_i g_i(y_i) | \alpha_i \in \mathbb{R}, g_i \in \mathcal{G}^1\}$. That is, a linear combination of functions of the individual y_i values. We can use these functions to deduce a sufficient, but not necessary, condition for observability.

Lemma 3.1: Assume that $|\mathcal{X}| = n$ and define the matrix

$$M = \begin{pmatrix} A & TA & \cdots & T^{n-1}A \end{pmatrix}$$

which is $n \times nK$ where $K = |\mathcal{Y}|$. If M is rank n, then the system is n-step observable. Furthermore, if M is not rank n, appending more blocks of the form $T^k A$ for $k \geq n$ will not increase the rank of M.

Proof: Beginning with (8), consider a restriction to $\mathcal{G}^n_{\text{LC}}$, that is, we require g to be of the form $g(y_{[1,n]}) = \sum_{i=1}^n g_i(y_i)$. Denote the $(nK) \times 1$ vector

$\alpha = (g_1(c_1), \dots, g_1(c_K), \dots, g_n(c_1), \dots, g_n(c_K))$. Then

$$\begin{aligned} f(x) &= \sum_{i=1}^n P^\mu(y_i \in \cdot | X_1 = x) \begin{pmatrix} g_i(c_1) \\ \vdots \\ g_i(c_K) \end{pmatrix} \\ &= \begin{pmatrix} QH_x & T(x| :)A & \cdots & T(x| :)T^{n-2}A \end{pmatrix} \alpha. \end{aligned}$$

We can then see that this matrix is the jth row of M when $x = a_j$; therefore, we have $\begin{pmatrix} f(a_1) \\ \vdots \\ f(a_n) \end{pmatrix} = \begin{pmatrix} A & TA & \cdots & T^{n-1}A \end{pmatrix}\alpha$. If M is rank n, then any function $f : \mathcal{X} \to \mathbb{R}$ can be expressed as a vector g put through matrix M and the system is observable.

Consider if M is not rank n and if we append another block $T^n A$ to M. By the Cayley–Hamilton theorem, T^n is a linear combination of lower powers of T, e.g., $T^n = \sum_{i=0}^n \alpha_i T^i$ for some coefficients α_i. Therefore, this additional block is a linear combination of the previous blocks, and adds no dimension to matrix M.

If the conditions of this lemma fail, i.e., M is not rank n, that means integrating g over the marginal measures cannot generate any f function. Yet, the product of the marginal measures is not the joint measure since $Y_i|X_1$ are not independent. Hence, working with the marginal measures only is not enough to determine observability as also noted in [54, Remark 13] in a slightly different setup.

Consider the following example. Let $\mathcal{X} = \{1, 2, 3, 4\}$ and $Y = 1_{X\le 2}$. This can be realized as

$$A = \begin{pmatrix} QH_1 \\ \vdots \\ QH_4 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

Consider the following transition kernel:

$$T = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

Notice that the odd and even rows are identical. If we consider the marginal measures of $Y_1|X_1, \ldots, Y_4|X_1$ we have the matrix

$$\begin{pmatrix} A & \cdots & T^3A \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0.75 & 0.25 & 0.5625 & 0.4375 & 0.609375 & 0.390625 \\ 0 & 1 & 0.50 & 0.50 & 0.6250 & 0.3750 & 0.593750 & 0.406250 \\ 1 & 0 & 0.75 & 0.25 & 0.5625 & 0.4375 & 0.609375 & 0.390625 \\ 1 & 0 & 0.50 & 0.50 & 0.6250 & 0.3750 & 0.593750 & 0.406250 \end{pmatrix}$$

which is only rank 3, not rank 4. Therefore, we cannot use the marginal measures to determine observability.

However, if we consider the joint measure of $(Y_1, Y_2)|X_1$, we have the matrix

$$A' = \begin{pmatrix} 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{pmatrix}$$

where row i is conditioned on $x = i$ and the columns are ordered in binary y_2y_1, e.g., $P(y_1 = 1, y_2 = 0|x_1 = 2)$ is row 2 column 3. This matrix is full rank; hence, the system is N-step observable with $N = 2$, even though the marginal measures failed to be full rank.

B. Compact State and Noise Spaces With Affine Observations

Consider $\mathcal{X}$ and $\mathcal{Z}$ as compact subsets of $\mathbb{R}$ and let $h(x, z) = a(z)x + b(z)$ for some functions a and$b$, where the image of $\mathcal{Z}$ under a and b is compact (this ensures that $\mathcal{Y}$ is compact). Note that for a fixed choice of z, this is an affine function of x. We will arrive at sufficient conditions for one-step observability. Since $\mathcal{X}$ is compact, the set of polynomials is dense in the set of continuous and bounded functions. Therefore, rather than working with a function $f \in C_b(\mathcal{X})$ without loss of generality we assume f is a polynomial. Let $\mathcal{M}_b(\mathbb{R})$ represent the measurable and bounded functions on the real line and consider the mapping

$$S : \mathcal{M}_b(\mathbb{R}) \to C_b(\mathbb{R}) \quad S(g)(\cdot) \mapsto \int_Z g(h(\cdot, z))Q(dz).$$

Let $\mathbb{R}[x]_n$ represent the polynomials on the real line up to degree n. Then, we have that $S(g)$ is invariant on $\mathbb{R}[x]_n$, that is, if g is polynomial of degree n, then $S(g)$ is a polynomial of degree n. Furthermore, the coefficients of $S(g)(x) = \sum_{i=0}^n \beta_i x^i$ can be related to the coefficients of $g(x) = \sum_{i=0}^n \alpha_i x^i$ by a linear transformation. Define $N(i, k) = \binom{i}{k} E(a(Z)^k b(Z)^{i-k})$, then by recursive application of binomial theorem we have

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} N(0,0) & N(1,0) & \cdots & N(n,0) \\ 0 & N(1,1) & \cdots & N(n,1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & N(n,n) \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$$

if we want to generate any polynomial, we require this matrix to be invertible, and since it is upper triangular this amounts to none of the diagonal entries being zero, that is, $E[a(z)^n] \neq 0\,\forall n \in \mathbb{N}$. Furthermore, we want g to be bounded so we must have $N(n, k) < \infty\,\forall n \in \mathbb{N}, k \in \{0, \ldots, i\}$.

B. Example

Consider $\mathcal{X} = [-10, 10]$, $\mathcal{Z} = [-1, 1]$, $Z \sim$ Uni$([-1, 1])$, and $y = z^2x + z$. We then have $\mathcal{Y} = [-11, 11]$. For any $n \in \mathbb{N}$, we have

$$E[a(z)^n] = \frac{1}{2}\int_{-1}^{1} z^{2n} dz = \frac{1}{2n+1} \neq 0$$

additionally, for any $n \in \mathbb{N}, k \in \{0, \ldots, n\}$, we have

$$N(n, k) = \binom{n}{k} E(a(z)^k b(z)^{n-k}) = \binom{n}{k} E(z^{n+k})$$
$$= \binom{n}{k} \frac{1}{n+k+1} < \infty.$$

C. Nonlinear Measurement Function

Consider $\mathcal{X}$ as a compact subset of $\mathbb{R}$, $\mathcal{Z} = \mathbb{R}$. Let $h(x, z) = 1_{x>z}x + 1_{x\le z}z$ and assume that Q admits a density with respect to Lebesgue. We have

$$\int_{\mathcal{Z}} g(h(x, z))Q(dz) = \int_{-\infty}^{x} g(x)q(z)dz + \int_{x}^{\infty} g(z)q(z)dz$$

again, we can approximate any continuous and bounded function f on $\mathcal{X}$ as polynomial, so we assume that f is differentiable. We have

$$f(x) = \int_{-\infty}^{x} g(x)q(z)dz + \int_{x}^{\infty} g(z)q(z)dz$$
$$f'(x) = g(x)q(x) + \int_{-\infty}^{x} g'(x)q(z)dz - g(x)q(x)$$
$$= g'(x)Q(Z \le x).$$

Since $\mathcal{X}$ is compact, there exists some $x_{\min} \in \mathbb{R}$ such that $x_{\min} < x \ \forall x \in \mathcal{X}$. We require for some $\epsilon > 0$ that $Q(Z < x_{\min}) > \epsilon$. This condition says every $x \in \mathcal{X}$ has some positive probability of being observed through $h(x,z)$ and we will not always get pure noise. Then, we have

$$g'(x) = 1_{\mathcal{X}}(x)\frac{f'(x)}{Q(Z \le x)}$$
$$g(x) = c + \int_{-\infty}^{x} 1_{\mathcal{X}}(u)\frac{f'(u)}{Q(Z \le u)}du$$

for some constant c. Therefore, we only need to define g over $\mathcal{X}$. Furthermore, we require g to be bounded, which is implied if g' is bounded since g is only defined over a compact space.

D. Local Observability for a Noncompact State Space

We now study a system that does not have a compact state signal space and satisfies the definitions of local predictability and local observability, so that we can apply Theorem 2.15. Consider the POMP with the following transition and measurement kernels:

$$X_{n+1} = X_n + N(1,1)$$
$$Y_n = \begin{cases} X_n + 1 & \text{w.p. } \frac{1}{2} \\ X_n - 1 & \text{w.p. } \frac{1}{2}. \end{cases}$$

We will first show that this system is locally predictable. Given an observation Y_{n-1}, it must be that $X_{n-1} = Y_{n-1} - 1$ or $Y_{n-1} + 1$; therefore, any filter at time $n-1$ will consist of two point masses at $Y_{n-1} - 1$ and $Y_{n-1} + 1$ with the probability of these two points dependent on the prior. Therefore, the predictor at time n will be a convex combination of Gaussian random variables $\alpha_n \mathcal{N}(Y_{n-1}, 1) + (1-\alpha_n)\mathcal{N}(Y_{n-1}+2, 1)$, where α_n is determined by the prior.

However, regardless of the value of α for any $\epsilon > 0$ have some compact set K_ϵ such that $\pi^{\nu}_{n-}(K_\epsilon + Y_{n-1}) > 1 - \epsilon$ for any choice of ν. Therefore, the system is locally predictable.

For local observability, assume that K is an interval $[-M, M]$ for some whole number $M > 0$ and pick a centering value a. Fix a continuous and bounded function f. We wish to demonstrate a function g that approximates f well over $K + a$ when integrated over the measurement channel. g must be bounded with a bound that does not depend on a. If we define $g(y)$ recursively as follows:

$$g(y) =$$
$$\begin{cases} 0 & y < -M + a + 1 \\ 2f(y-1) & y \in [-M+a+1, -M+a+3) \\ 2f(y-1) - g(y-2) & y \in [-M+a+3, M+a+1] \\ -g(y-2) & y > M+a+1 \end{cases}$$

g is akin to a telescoping sum in that it cancels out its own previous values. We have

$$\int g(h(x,z))Q(dz) = \frac{1}{2}(g(x+1) + g(x-1)).$$

For $x < -M + a$, we have $x - 1 < x + 1 < -M + a + 1$; hence, $g(x-1) = g(x+1) = 0$. For $x \in [-M+a, -M+a+2)$, we have $g(x+1) = 2f(x+1-1) = 2f(x)$ while $g(x-1) = 0$. Then, for $x \in [-M+a+2, M+a]$, we have

$$g(x+1) = 2f(x+1-1) - g(x+1-2) = 2f(x) - g(x-1)$$

which will cancel with the other $g(x-1)$ term, hence the telescoping. For $x > M + a$, we have $g(x+1) = -g(x+1-2) = -g(x-1)$; hence, it will cancel with the previous value.

In each iteration of telescoping, $\|g\|_\infty$ increases by at most $2\|f\|_\infty$, there are $2M$ iterations of telescoping so the overall bound on $\|g\|_\infty$ is $4M\|f\|_\infty$. Therefore, we have

$$\|g\|_\infty \le 4M\|f\|_\infty$$
$$\int g(h(x,z)Q(dz) = f(x) \qquad x \in [-M+a, M+a]$$
$$\left|\int g(h(x,z)Q(dz)\right| = 0 \qquad x \notin [-M+a, M+a]$$

this proves local observability.

IV. Proofs

A. Observability: Proof of Theorem 2.7

Lemma 4.1: Let g be a bounded and measurable function on $(\mathcal{Y}^{k+1}, \mathcal{B}(\mathcal{Y}^{k+1}))$. For any initial prior μ, we have

$$\int_{\mathcal{Y}^{k+1}} g(y_{[n,n+k]})P^{\mu}(dy_{[n,n+k]}|Y_{[0,n-1]})$$
$$= \int_{\mathcal{X}}\int_{\mathcal{Y}^{k+1}} g(y_{[n,n+k]})P(dy_{[n,n+k]}|X_n = x_n)\pi^{\mu}_{n-}(dx_n). \tag{9}$$

Proof:

$$\int_{\mathcal{Y}^{k+1}} g(y_{[n,n+k]})P^{\mu}(dy_{[n,n+k]}|Y_{[0,n-1]})$$
$$= \int_{\mathcal{Y}^{k+1}\times\mathcal{X}} g(y_{[n,n+k]})P^{\mu}(d(y_{[n,n+k]}, x_n)|Y_{[0,n-1]})$$

we then apply the chain rule for conditional probability measures and we have

$$\int_{\mathcal{X}}\int_{\mathcal{Y}^{k+1}} g(y_{[n,n+k]})P^{\mu}(dy_{[n,n+k]}|x_n, Y_{[0,n-1]})\pi^{\mu}_{n-}(dx_n).$$

Since $\{(X_n, Y_n)\}_{n=0}^{\infty}$ is a Markov chain, $Y_{[n,n+k]}$ is conditionally independent of $Y_{[0,n-1]}$ given X_n. In addition, the prior does not determine the conditional measure; therefore, we have

$$\int_{\mathcal{X}}\int_{\mathcal{Y}^{k+1}} g(y_{[n,n+k]})P(dy_{[n,n+k]}|x_n)\pi^{\mu}_{n-}(dx_n)$$

where we do not include a prior in the superscript of the conditional measure, since the conditional measure is the same regardless of the prior.

Corollary 4.2: Let g be a bounded and measurable function on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. For any prior μ, we have

$$\int_{\mathcal{Y}} g(y_n) P^{\mu}(dy_n | X_n = x) = \int_{\mathcal{Z}} g(h_x(z)) Q(dz). \tag{10}$$

Proof: Z is a random variable on the probability space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), Q)$ and Y_n exists on the measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. Then, for every fixed choice of $X_n = x$, we have that Y_n is a fixed function of Z, that is $Y_n = h_x(Z)$. For any set $A \in \mathcal{B}(\mathcal{Y})$, we have $P^{\mu}(Y_n \in A | X_n = x) = Q(h_x^{-1}(A))$. Yet this means that $P^{\mu}(Y_n \in \cdot | X_n = x)$ is exactly the pushforward measure of Q under the mapping h_x, call this measure $h_x Q(A) = Q(h_x^{-1}(A))$. We then have

$$\int_{\mathcal{Y}} g(y) h_x Q(dy)) = \int_{\mathcal{Z}} g(h_x(z)) Q(dz).$$

Notice that the inner integral in the right-hand side (RHS) of (9) is a function of x. The left-hand side (LHS) is then the term considered in the total variation merging of the predictive measures of the measurement sequences, whereas the RHS is the term considered in the weak merging of the one-step predictor. We can then leverage the Blackwell and Dubins theorem to arrive at a sufficient condition for weak merging of the one-step predictor. Theorem 2.7 is closely related to [56, Prop. 3.11] and its proof is in essence a sufficient condition for uniform observability (of the predictor).

Proof of Theorem 2.7

Fix any $f \in C_b(\mathcal{X})$ and $\epsilon > 0$. We wish to show that $\exists N$ such that $\forall n > N$

$$\left| \int f d\pi_{n-}^{\mu} - \int f d\pi_{n-}^{\nu} \right| < \epsilon.$$

By observability for the fixed f, (5) holds for some $N' + 1$. Therefore, we can find some g with $\|g\|_\infty < \infty$ such that

$$\tilde{f}(x) = \int_{\mathcal{Y}^{N'+1}} g(y_{[1,1+N']}) P(dy_{[1,1+N']} | X_1 = x)$$

and $\|f - \tilde{f}\|_\infty < \frac{\epsilon}{3}$. Conditioned on the value of $X_n = x$ and since the noise is i.i.d., the conditional channel $Y_{[n,n+N']} | X_n$ is time invariant, so it holds that

$$\tilde{f}(x) = \int_{\mathcal{Y}^{N'+1}} g(y_{[n,n+N']}) P(dy_{[n,n+N']} | X_n = x)$$

is the same regardless of the choice of n. Then, we have

$$\begin{aligned} &\left| \int f d\pi_{n-}^{\mu} - \int f d\pi_{n-}^{\nu} \right| \\ &\leq \left| \int \tilde{f} d\pi_{n-}^{\mu} - \int \tilde{f} d\pi_{n-}^{\nu} \right| + \left| \int (f - \tilde{f}) d\pi_{n-}^{\mu} \right| \\ &\quad + \left| \int (f - \tilde{f}) d\pi_{n-}^{\nu} \right|. \end{aligned} \tag{11}$$

Now, by assumption, $\|f - \tilde{f}\|_\infty < \frac{\epsilon}{3}$; therefore, the last two terms are less than $\frac{2}{3}\epsilon$. We then apply Lemma 4.1 and we have

$$\begin{aligned} &\left| \int \tilde{f} d\pi_{n-}^{\mu} - \int \tilde{f} d\pi_{n-}^{\nu} \right| + \frac{2}{3}\epsilon \\ &= \left| \int_{\mathcal{Y}^{N'+1}} g(y_{[n,n+N']}) P^{\mu}(dy_{[n,n+N']} \right| Y_{[0,n-1]}) \\ &\quad - \int_{\mathcal{Y}^{N'+1}} g(y_{[n,n+N']}) P^{\nu}(dy_{[n,n+N']} | Y_{[0,n-1]})| + \frac{2}{3}\epsilon. \end{aligned}$$

By Assumption 6, we have $P^{\mu}(Y_{[0,\infty)} \in \cdot) \ll P^{\nu}(Y_{[0,\infty)} \in \cdot)$. Then, via a classic result by Blackwell and Dubins [7], we have that $P^{\mu}(Y_{[n,n+N']} \in \cdot | Y_{[0,n-1]})$ and $P^{\nu}(Y_{[n,n+N']} \in \cdot | Y_{[0,n-1]})$ merge in total variation P^{μ} a.s. as $n \to \infty$. Define $\tilde{g} = \frac{g}{\|g\|_\infty}$. Then, $\exists\, N \in \mathbb{N}$ such that $\forall n > N$

$$\begin{aligned} &\left| \int_{\mathcal{Y}^{N'+1}} \tilde{g}(y_{[n,n+N']}) P^{\mu}(dy_{[n,n+N']} \right| Y_{[0,n-1]}) \\ &\quad - \int_{\mathcal{Y}^{N'+1}} \tilde{g}(y_{[n,n+N']}) P^{\nu}(dy_{[n,n+N']} | Y_{[0,n-1]})| < \frac{\epsilon}{3\|g\|_\infty} \end{aligned}$$

we then have

$$\begin{aligned} &\left| \int_{\mathcal{Y}^{N'+1}} g(y_{[n,n+N']}) P^{\mu}(dy_{[n,n+N']} \right| Y_{[0,n-1]}) \\ &\quad - \int_{\mathcal{Y}^{N'+1}} g(y_{[n,n+N']}) P^{\nu}(dy_{[n,n+N']} | Y_{[0,n-1]})| + \frac{2}{3}\epsilon \\ &\leq \|g\|_\infty \frac{\epsilon}{3\|g\|_\infty} + \frac{2}{3}\epsilon = \epsilon \end{aligned}$$

therefore, since f and ϵ are arbitrary, we have for any $f \in C_b(S)$: $\lim_{n\to\infty} |\int f d\pi_{n-}^{\mu} - \int f d\pi_{n-}^{\nu}| = 0$, which means π_{n-}^{μ} and π_{n-}^{ν} merge weakly.

B. Weak Filter Stability: Proof of Theorem 2.9

Proof of Theorem 2.9

Begin by assuming that the predictor merges weakly a.s. As is argued in [31], one can view the filter π_n^{μ} as a function of π_{n-1}^{μ} (the previous filter) and the current observation $Y_n = y_n$, that is, $\pi_n^{\mu} = F(\pi_{n-1}^{\mu}, y_n)$. Picking any continuous and bounded function f, we have

$$\begin{aligned} &E^{\mu}\left[\left| \int_{\mathcal{X}} f(x) \pi_n^{\mu}(dx) - \int_{\mathcal{X}} f(x) \pi_n^{\nu}(dx) \right|\right] \\ &= E^{\mu}\left[E^{\mu}\left[\left| \int_{\mathcal{X}} f(x) F(\pi_{n-1}^{\mu}, y_n)(dx) \right.\right.\right. \\ &\quad \left.\left.\left. - \int_{X} f(x) F(\pi_{n-1}^{\nu}, y_n)(dx) \right| \Big| Y_{[0,n-1]} \right]\right]. \end{aligned} \tag{12}$$

Now, define the set $I^{+}(y_{[0,n-1]}) \subset \mathcal{Y}$ as

$$\begin{aligned} I^{+}(y_{[0,n-1]}) = \Big\{ y_n \in \mathbb{Y} \Big| &\int_{\mathcal{X}} f(x) F(\pi_{n-1}^{\mu}, y_n)(dx) \\ &> \int_{X} f(x) F(\pi_{n-1}^{\nu}, y_n)(dx) \Big\} \end{aligned}$$

where the argument $y_{[0,n-1]}$ is the sequence on which the previous filters π_{n-1}^{μ} and π_{n-1}^{ν} are realized. Define the complement of this set as $I^{-}(y_{[0,n-1]})$. Then, for every fixed realization $y_{[0,n-1]}$, we can break the inner expectation in (12) (which is an integral) into two parts and follow the analysis in [31, eq. 4] together with [8, Th. 8.6.2] to arrive at the conclusion.

C. Local Observability: Proof of Theorems 2.12 and 2.15

The idea of local observability is the shift some of the burden of approximating the signal f. When we work with a function

$$\tilde{f}(x) = \int_{\mathcal{Y}^{N'+1}} g(y_{[n,n+N']}) P(dy_{[n,n+N']} | X_n = x)$$

the result is the terms seen in (11). The first term is dealt with by the Blackwell and Dubins theorem, so we must make sure the second and third term can be made arbitrarily small. For any set K, we can write

$$\left| \int (f - \tilde{f}) d\pi_{n-}^{\nu} \right| \leq \sup_{x \in K} |f(x) - \tilde{f}(x)| \pi_{n-}^{\nu}(K) + \sup_{x \notin K} |f(x) - \tilde{f}(x)| \pi_{n-}^{\mu}(K^C)$$

in the previous result we bounded this by simply approximating f well over the whole space. Instead, we can choose a K where $\tilde{f}$ approximates f well over K and $\pi_{n-}^{\nu}(K^C)$ makes the other term arbitrarily small. Furthermore, by taking advantage of the full supremum of total variation, we can work with a series of uniformly bounded functions $\tilde{f}_n$ and shifting sets K_n that change with n.

Proof of Theorem 2.12: Pick any continuous and bounded function f and any $\epsilon > 0$. Fix any sequence of observations $y_{[0,\infty)}$, where the predictors π_{n-}^{μ} and π_{n-}^{ν} are well defined and maintain this sequence for the rest of the proof. Then, consider

$$\lim_{n \to \infty} \left| \int_{\mathcal{X}} f(x) \pi_{n-}^{\mu}(dx) - \int_{\mathcal{X}} f(x) \pi_{n-}^{\nu}(dx) \right|.$$

For any function series of functions $\tilde{f}_n$ of x, we have an upper bound

$$\lim_{n \to \infty} \left| \int_{\mathcal{X}} \tilde{f}_n(x) \pi_{n-}^{\mu}(dx) - \int_{\mathcal{X}} \tilde{f}_n(x) \pi_{n-}^{\nu}(dx) \right| + \left| \int_{\mathcal{X}} (f - \tilde{f}_n)(x) \pi_{n-}^{\mu}(dx) \right| + \left| \int_{\mathcal{X}} (f - \tilde{f}_n)(x) \pi_{n-}^{\nu}(dx) \right|.$$

By assumption of K local predictability, we have a compact sets $K_n = K + a_n$, where $\pi_{n-}^{\mu}(K_n) = 1$ for every $\mu \ll \nu$ and every n.

By K local observability, we can find a uniformly bounded series of functions $g_n \leq M$, where

$$\tilde{f}_n(x) = \int_{\mathcal{Z}} g_n(h(x,z)) Q(dz)$$

$$\sup_{x \in K_n} \left| f(x) - \tilde{f}_n(x) \right| \leq \frac{\epsilon}{3}$$

then for the two approximation terms, we have

$$\begin{aligned} &\left| \int_{\mathcal{X}} (f - \tilde{f}_n)(x) \pi_{n-}^{\nu}(dx) \right| \\ &\leq \sup_{x \in K_n} |f(x) - \tilde{f}_n(x)| \pi_{n-}^{\nu}(K_n) \\ &\quad + \sup_{x \notin K_n} |f(x) - \tilde{f}_n(x)| \pi_{n-}^{\mu}(K_n^C) \\ &\leq \frac{\epsilon}{3} \end{aligned}$$

we then have

$$\begin{aligned} &\lim_{n \to \infty} \left| \int_{\mathcal{X}} \tilde{f}_n(x) \pi_{n-}^{\mu}(dx) - \int_{\mathcal{X}} \tilde{f}_n(x) \pi_{n-}^{\nu}(dx) \right| + \frac{2}{3}\epsilon \\ &= \lim_{n \to \infty} \left| \int_{\mathcal{Y}} g_n(y_n) P^{\mu}(dy_n | y_{[0,n-1]}) \right. \\ &\quad \left. - \int_{\mathcal{Y}} g_n(y_n) P^{\nu}(dy_n | y_{0,n-1]}) \right| + \frac{2}{3}\epsilon \end{aligned}$$

we must appeal to the full uniform bound of the Blackwell and Dubins theorem, which was not required in the proof of Theorem 2.7. The full statement of the Blackwell and Dubins theorem tells us that

$$\lim_{n \to \infty} \sup_{\|g\| \leq 1} \left| \int_{\mathcal{Y}} g(y_n) P^{\mu}(dy_n | y_{[0,n-1]}) - \int_{\mathcal{Y}} g(y_n) P^{\nu}(dy_n | y_{0,n-1]}) \right| = 0 \tag{13}$$

where the supremum is taken over measurable functions g. Thus, for any fixed measurable and bounded function g, we have that

$$\left| \int_{\mathcal{Y}} g(y_n) P^{\mu}(dy_n | y_{[0,n-1]}) - \int_{\mathcal{Y}} g(y_n) P^{\nu}(dy_n | y_{0,n-1]}) \right|$$

converges to 0 as $n \to \infty$; this was the form of the statement utilized in the proof of Theorem 2.7. However, if we have a sequence of measurable functions g_n with a uniform bound, $g_n \leq M \, \forall \, n \in \mathbb{N}$, then the supremum in (13) allows us to make a uniform claim about the convergence to zero of the sequence

$$\left| \int_{\mathcal{Y}} g_n(y_n) P^{\mu}(dy_n | y_{[0,n-1]}) - \int_{\mathcal{Y}} g_n(y_n) P^{\nu}(dy_n | y_{0,n-1]}) \right|$$

and this completes the proof.

Proof of Theorem 2.15: Fix any f and any ϵ. We begin from the upper bound used previously

$$\lim_{n \to \infty} \left| \int_{\mathcal{X}} \tilde{f}_n(x) \pi_{n-}^{\mu}(dx) - \int_{\mathcal{X}} \tilde{f}_n(x) \pi_{n-}^{\nu}(dx) \right| + \left| \int_{\mathcal{X}} (f - \tilde{f}_n)(x) \pi_{n-}^{\mu}(dx) \right| + \left| \int_{\mathcal{X}} (f - \tilde{f}_n)(x) \pi_{n-}^{\nu}(dx) \right|$$

for some series of functions $\tilde{f}_n$.

By local predictability, the shifted predictors are a tight family. Therefore, for any ϵ', we have a series of compact sets $K_n = K' + a_n$ such that $\pi_{n-}^{\nu}(K_n) \geq 1 - \epsilon'$ for any $\mu \ll \nu$ and any n.

The proof then proceeds similarly to that of Theorem 2.12.

D. Predictor Merging in Total Variation: Proof of Theorem 2.18

We now extend our results from weak merging to total variation. We first state the following supporting results.

Lemma 4.3: The (measurement update) map

$$(\pi_{n_-}, y) \mapsto \pi_n \quad : \quad \pi_n(\cdot) := E_{\pi_{n_-}}[1_{X_n \in \cdot} | Y_n = y]$$

which maps from $\mathcal{P}(\mathcal{X}) \times \mathbb{Y}$ to $\mathcal{P}(\mathcal{X})$ is weakly continuous in π_{n_-} for almost every y, provided that $g(x,y)$ is positive, bounded, and continuous in x for every fixed y.

Proof: Consider a continuous and bounded f and let $\pi^m_{n_-} \to \pi_{n_-}$ weakly. Then

$$E_{\pi^m_{n_-}}[f(x_n)|Y_n = y_n] = \int f(x_n) \frac{g(x_n, y_n)\pi^m_{n_-}(dx_n)}{\int_{\mathcal{X}} g(x_n, y_n)\pi^m_{n_-}(dx_n)}$$
$$= \frac{\int f(x_n) g(x_n, y_n)\pi^m_{n_-}(dx_n)}{\int_{\mathcal{X}} g(x_n, y_n)\pi^m_{n_-}(dx_n)}.$$

Since $g(\cdot, y_n)$ is bounded and continuous, both the numerator and the denominator converge. $\square$

Lemma 4.4: Let $T(dx_1|x) = t(x_1, x)\phi(dx_1)$ where t is continuous in x for every x_1. Then, the (time-update) map:

$$(\pi_n) \mapsto \pi_{n+1_-} \quad : \quad \pi_{n+1_-}(\cdot) := \int T(\cdot|x_n)\pi_n(dx_n)$$

which maps from $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{X})$ is so that if $\pi^\nu_n \to \pi^\mu_n$ weakly then $\pi^\nu_{n+1_-} \to \pi^\mu_{n+1_-}$ in total variation.

Proof: We will build on Scheffé's Lemma [5]. For every given history, we have

$$\pi^\nu_{n+1_-}(dx_{n+1}) = \int T(dx_{n+1}|x_n)\pi^\nu_n(dx_n).$$

Now, $\int T(dx_{n+1}|x_n)$ is so that

$$\int t(x_{n+1}, x_n)\phi(dx_{n+1})\pi^m_n(dx_n)$$
$$\to \int t(x_{n+1}, x_n)\phi(dx_{n+1})\pi_n(dx_n)$$

in total variation since for every fixed z, the Radon–Nikodym derivative (density) with respect to ϕ

$$\frac{\int t(x_{n+1}, x_n)\phi(\cdot)\pi^m_n(dx_n)}{d\phi}(z) = \int t(z, x_n)\pi^m_n(dx_n)$$

satisfies pointwise convergence

$$\int t(z, x_n)\pi^\nu_n(dx_n) \to \int t(z, x_n)\pi^\mu_n(dx_n)$$

and Scheffé's lemma implies that convergence is in total variation. Now, we can apply the result to the sequence π^ν_n converging to π^μ_n.

Proof: Proof of Theorem 2.18 (i)

Under Assumption 2.16, the proof follows from Lemmas 4.3 and 4.4. While in Lemmas 4.3 and 4.4 we consider convergence (and not merging), we note that the proof of Lemma 4.3 also implies weak merging of the posteriors as the priors weakly merge, and by considering the signed measure $\pi^{\nu,\gamma}_n - \pi^{\mu,\gamma}_n$ in the proof of Lemma 4.4, total variation merging is a result of a generalized Scheffé's lemma [8, Th. 2.8.9].

Lemma 4.5: Let $\exists$ some measure $\bar{\mu}$ such that $T(\cdot|x) \ll \bar{\mu}$ for every $x \in \mathcal{X}$. Then, we have that $\pi^\mu_{n-}, \pi^\nu_{n-} \ll \bar{\mu}$ for every $n \in \mathbb{N}$.

Proof: For all $n \geq 1$, we have

$$\pi^\mu_{n-}(A) = \int_{\mathcal{X}} T(A|x)\pi^\mu_{n-1}(dx)$$
$$= \int_{\mathcal{X}}\int_A \frac{dT(\cdot|x)}{d\bar{\mu}}(a)\bar{\mu}(da)\pi^\mu_{n-1}(dx)$$
$$= \int_A \left(\int_{\mathcal{X}} \frac{dT(\cdot|x)}{d\bar{\mu}}(a)\pi^\mu_{n-1}(dx)\right)\bar{\mu}(da)$$

where we have applied Fubini's theorem in the final equality. Therefore, π^μ_{n-} is absolutely continuous with respect to $\bar{\mu}$ for every $n \geq 1$.

Lemma 4.6: Let Assumption 2.17 hold and let f^μ_{n-} denote the density function of π^μ_{n-}. Fix any sequence of measurements $y_{[0,\infty)}$ and denote the collection of probability density functions $\mathscr{F}^\mu = \{f^\mu_{n-}|n \in \mathbb{N}\}, \mathscr{F}^\nu = \{f^\nu_{n-}|n \in \mathbb{N}\}$. Then, $\mathscr{F}^\mu$ and $\mathscr{F}^\nu$ are uniformly bounded equicontinuous families.

Proof: As we see from Lemma 4.5

$$f^\mu_{n-}(x_n) = \frac{d\pi^\mu_{n-}}{d\phi}(x_n) = \int_{\mathcal{X}} t(x_n|x_{n-1})\pi^\mu_{n-1}(dx_{n-1})$$

where $t(\cdot|x)$ is the Radon–Nikodym derivative of $T(\cdot|x)$ with respect to our dominating measure ϕ and $d(\cdot,\cdot)$ will represent the metric on $\mathcal{X}$ (recall $\mathcal{X}$ is a complete, separable, metric space). We require $\forall \epsilon > 0$, $x^* \in \mathcal{X} \,\exists\, \delta > 0$ such that $\forall\, d(x, x^*) < \delta$ $\forall n \in \mathbb{N}$ we have $|f^\mu_{n-}(x) - f^\mu_{n-}(x^*)| < \epsilon$. By Assumption 2.17, clearly f^μ_{n-} is uniformly bounded since t is uniformly bounded. Then, for any $\epsilon > 0$ $\forall x^* \in \mathcal{X}$, we can find a $\delta > 0$ such that $\forall x_1 \in \mathcal{X}, |t(x_2|x_1) - t(x^*|x_1)| < \epsilon$ when $d(x_2, x^*) < \delta$. Now, assume $d(x_2, x^*) < \delta$, we have

$$|f^\mu_{n-}(x_2) - f^\mu_{n-}(x^*)| = \left|\int_{\mathcal{X}} t(x_2|x_1) - t(x^*|x_1) d\pi^\mu_{n-}(dx_1)\right|$$
$$\leq \int_{\mathcal{X}} |t(x_2|x_1) - t(x^*|x_1)| d\pi^\mu_{n-}(x_1) \leq \epsilon$$

which proves that $\mathscr{F}^\mu$ and $\mathscr{F}^\nu$ are uniformly bounded and equicontinuous families.

Proof of Theorem 2.18 (ii): By assumption, we have weak stability of the predictor P^μ a.s. Then, there exists a set of measure sequences $B \subset \mathcal{Y}^{\mathbb{Z}_+}$ with $P^\mu(B) = 1$. For each measurement sequence $y_{[0,\infty]} \in B$, we have that the predictor realizations π^μ_{n-} and π^ν_{n-} merge in the weak sense. We will choose a general measurement sequence $y_{[0,\infty]} \in B$ and fix this sequence for the rest of the proof. Via Lemmas 4.5 and 4.6, $\mathscr{F}^\mu$ and $\mathscr{F}^\nu$ are uniformly bounded and equicontinuous families. Let $\mathscr{F}^{\mu-\nu} = \{f_n | f_n = f^\mu_{n-} - f^\nu_{n-}]\}$, then the sequence $\{f_n\}_{n=1}^\infty$ is a uniformly bounded and equicontinuous class of integrable functions. As in the proof of [40, Lemma 2], now pick a sequence of compact sets $K_j \subset \mathcal{X}$ such that $K_j \subset K_{j+1}$. By the Arzela–Ascoli theorem [49], for any subsequence, we can find further subsequences $f_{n^j_k}$ such that

$$\lim_{k\to\infty} \sup_{x\in K_j} |f_{n^j_k}(x) - f^j(x)| = 0$$

for some continuous function $f^j : K_j \to [0, \infty)$. Via the K_j being nested, we can have $\{f_{n^{j+1}_k}\}$ be a subsequence of $\{f_{n^j_k}\}$, and therefore $f^{j+1} = f^j$ over K_j. Then, define the function $\tilde{f}$ on $\mathcal{X}$ by $\tilde{f}(x) = f^j(x), x \in K_j$. Using Cantor's diagonal method, we can find an increasing sequence of integers $\{m_i\}$, which is a subsequence of $\{n^j_k\}$ for every j. Therefore,

$$\lim_{i\to\infty} f_{m_i}(x) = \tilde{f}(x) \quad \forall x \in \mathcal{X}$$

and the convergence is uniform over each K_j and $\tilde{f}$ is continuous. Now, f_{m_i} converges weakly to the zero measure by assumption, and via uniform convergence for any Borel set $\mathcal{B}$, we have

$$\int_{\mathcal{B}} f_{m_i}(x)dx \to \int_{\mathcal{B}} \tilde{f}(x)dx$$

i.e., setwise convergence. Yet this implies weak convergence, so $\tilde{f} = 0$ almost everywhere, yet $\tilde{f}$ is continuous so it is 0 everywhere. Now, via the Prokhorov theorem (see [8, Th. 8.6.2]), we have that $\mathscr{F}^{\mu-\nu}$ is a tight family. Therefore, for every $\epsilon > 0$, we can find a compact set K_ϵ such that

$$|\pi^\mu_{n-} - \pi^\nu_{n-}|(\mathcal{X} \setminus K_\epsilon) < \epsilon \quad \forall n \in \mathbb{N}.$$

Then, we have

$$\begin{aligned}\lim_{i\to\infty} \|\pi^\mu_{m_i-} - \pi^\nu_{m_i-}\|_{TV} &\leq \lim_{i\to\infty} |\pi^\mu_{m_i-} - \pi^\nu_{m_i-}|(\mathcal{X} \setminus K_\epsilon) \\ &\quad + |\pi^\mu_{m_i-} - \pi^\nu_{m_i-}|(K_\epsilon) \\ &\leq \lim_{i\to\infty} \sup_{\|g\|_\infty \leq 1} \left| \int_{K_\epsilon} g(x) f_{m_i}(x)dx \right| + \epsilon \\ &\leq \lim_{i\to\infty} \sup_{\|g\|_\infty \leq 1} \left| \int_{K_\epsilon} g(x)(\tilde{f} - f_{m_i})(x)dx \right| \\ &\quad + \left| \int_{K_\epsilon} g(x)\tilde{f}(x)dx \right| + \epsilon \\ &\leq \lim_{i\to\infty} \|\tilde{f} - f_{m_i}\|_\infty \phi(K_\epsilon) + \epsilon\end{aligned}$$

since we have already argued $\tilde{f} = 0$. Now, over the compact set K_ϵ, f_{m_i} converges to $\tilde{f}$ uniformly, therefore $\exists N$ such that $\forall k > N$, $\|\tilde{f} - f_{n_k}\|_\infty < \frac{\epsilon}{\phi(K_\epsilon)}$. We then conclude that

$$\lim_{i\to\infty} \|\pi^\mu_{m_i-} - \pi^\nu_{m_i-}\|_{TV} = 0.$$

Thus, for every subsequence of $\{f_n\}_{n=1}^\infty$, we can find a subsequence that converges in total variation, which implies that the original sequence converges in total variation.

E. Filter Merging in Total Variation: Proof of Theorem 2.19

For completeness, in the Appendix, some supporting results are presented.

Proof of Theorem 2.19: The sigma fields $\mathcal{F}^{\mathcal{X}}_{n,\infty} \vee \mathcal{F}^{\mathcal{Y}}_{0,\infty}$ are a decreasing sequence, that is, $\mathcal{F}^{\mathcal{X}}_{n+1,\infty} \vee \mathcal{F}^{\mathcal{Y}}_{0,\infty} \subset \mathcal{F}^{\mathcal{X}}_{n,\infty} \vee \mathcal{F}^{\mathcal{Y}}_{0,\infty}$. Therefore, when we take their intersection, removing the first or largest sigma field $\mathcal{F}^{\mathcal{X}}_{0,\infty} \vee \mathcal{F}^{\mathcal{Y}}_{0,\infty}$ from the intersection of a deceasing set of sigma fields does not change the overall intersection. From Lemmas A.5 and A.8, it is clear that the two conditions for merging in total variation in expectation are equivalent since the sigma fields on the LHS of (16) and (19) are equal.

We have now established that the filter merges in total variation in expectation, but we would like to extend this result to a.s. By a simple application of Fatou's lemma, we can argue the liminf of the total variation of the filter is zero P^μ a.s. Hence, if the limit exists, it must be zero, yet it is not immediate that the limit will exist. This leads to the following.

Theorem 4.7 (see [53, p. 572]): Assume that the filter is stable in total variation in expectation. Then, the filter is stable in total variation P^μ a.s.

F. Relative Entropy Merging: Proof of Theorem 2.20

We will now show that the relative entropy merging of the filter is essentially equivalent to merging in total variation in expectation. Via Lemmas A.4 and A.6, it is clear that the filter and predictor admit Radon–Nikodym derivatives. Therefore, working with $D(\pi^\mu_n\|\pi^\nu_n)$ and $D(\pi^\mu_{n-}\|\pi^\nu_{n-})$ is well defined. A well-known result for relative entropy is the chain rule [25, Th. 5.3.1].

Lemma 4.8: For joint measures P and Q on random variables X and Y, we have

$$\begin{aligned}D(P(X,Y)\|Q(X,Y)) &= D(P(X)\|Q(X)) \\ &\quad + D(P(Y|X)\|Q(Y|X)).\end{aligned}$$

Note for two sigma fields $\mathcal{F}$ and $\mathcal{G}$ and two joint measures P and Q on $\mathcal{F} \vee \mathcal{G}$, one could also express this relationship as

$$D(P|_{\mathcal{F}\vee\mathcal{G}}\|Q|_{\mathcal{F}\vee\mathcal{G}}) = D(P|_{\mathcal{F}}\|Q|_{\mathcal{F}}) + D(P|_{\mathcal{G}}|\mathcal{F}\|Q|_{\mathcal{G}}|\mathcal{F}). \tag{14}$$

Proof of Theorem 2.20: First assume the filter is stable in relative entropy. Since the square root function is continuous and convex, we have

$$\begin{aligned}0 &= \lim_{n\to\infty} \sqrt{\frac{2}{\log(e)} E^\mu[D(\pi^\mu_n\|\pi^\nu_n)]} \\ &\geq \lim_{n\to\infty} E^\mu\left[\sqrt{\frac{2}{\log(e)} D(\pi^\mu_n\|\pi^\nu_n)}\right]\end{aligned}$$

where we have applied Jensen's inequality. We then apply Pinsker's inequality, and we have $\lim_{n\to\infty} E^\mu[\|\pi^\mu_n - \pi^\nu_n\|_{TV}] = 0$.

For the converse direction, by chain rule (14), it is clear that

$$\begin{aligned}E^\mu[D(\pi^\mu_n\|\pi^\nu_n)] &= D(P^\mu|_{\mathcal{F}^{\mathcal{X}}_n}|\mathcal{F}^{\mathcal{Y}}_{0,n}\|P^\nu|_{\mathcal{F}^{\mathcal{X}}_n}|\mathcal{F}^{\mathcal{Y}}_{0,n}) \\ &= D(P^\mu|_{\mathcal{F}^{\mathcal{X}}_n\vee\mathcal{F}^{\mathcal{Y}}_{0,n}}\|P^\nu|_{\mathcal{F}^{\mathcal{X}}_n\vee\mathcal{F}^{\mathcal{Y}}_{0,n}}) - D(P^\mu|_{\mathcal{F}^{\mathcal{Y}}_{0,n}}\|P^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,n}})\end{aligned}$$

by the Markov property, we have that $X_{[0,n-1]}, Y_{[0,n-1]}$ and $X_{[n+1,\infty)}, Y_{[n+1,\infty)}$ are conditionally independent given X_n and Y_n; therefore, we have

$$\begin{aligned}&D(P^\mu|_{\mathcal{F}^{\mathcal{X}}_n\vee\mathcal{F}^{\mathcal{Y}}_{0,n}}\|P^\nu|_{\mathcal{F}^{\mathcal{X}}_n\vee\mathcal{F}^{\mathcal{Y}}_{0,n}}) \\ &\quad = D(P^\mu|_{\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}\|P^\nu|_{\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}).\end{aligned}$$

Then, $\mathcal{F}^{\mathcal{X}}_{n,\infty} \vee \mathcal{F}^{\mathcal{Y}}_{0,\infty}$ is a decreasing sequence of sigma fields. By [4, Th. 2], we have that if the relative entropy is ever finite, the limit of the relative entropy restricted to these sigma fields is the relative entropy restricted to the intersection of the decreasing fields, that is,

$$\begin{aligned}&\lim_{n\to\infty} D(P^\mu|_{\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}\|P^\nu|_{\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}) \\ &\quad = D(P^\mu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}\|P^\nu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}).\end{aligned}$$

Likewise, $\mathcal{F}^{\mathcal{Y}}_{0,n}$ is an increasing sequence of sigma fields, therefore by [4, Th. 3], we have that if the relative entropy is ever finite, the relative entropy restricted to these sigma fields is the relative entropy over the limit field, that is,

$$\lim_{n\to\infty} D(P^\mu|_{\mathcal{F}^{\mathcal{Y}}_{0,n}}\|P^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,n}}) = D(P^\mu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}\|P^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}).$$

Therefore,

$$\begin{aligned}&\lim_{n\to\infty} E^\mu[D(\pi_n^\mu\|\pi_n^\nu)]\\&\quad= D(P^\mu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}\|P^\nu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}})\\&\qquad - D(P^\mu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}\|P^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}).\end{aligned}$$

By Lemma 1.1, we have

$$\frac{dP^\mu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}}{dP^\nu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}} = E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle|\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}\right] = f_1$$

$$\frac{dP^\mu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}}{dP^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}} = E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle|\mathcal{F}^{\mathcal{Y}}_{0,\infty}\right] = f_2.$$

Note that f_1 is $\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}$ measurable, while f_2 is $\mathcal{F}^{\mathcal{Y}}_{0,\infty}$ measurable, and $\mathcal{F}^{\mathcal{Y}}_{0,\infty}\subset\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}$. By Lemma A.5, we have that if the filter merges in total variation in expectation, then for a set of state and observation sequences $\omega = (x_i, y_i)_{i=0}^\infty \in A \subset \mathcal{F}^{\mathcal{X}}_{0,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}$ with $P^\nu(A) = 1$, we have $f_1(\omega) = f_2(\omega)$. Yet this then means over the set A of P^ν measure 1, $f_1 = f_2$ is $\mathcal{F}^{\mathcal{Y}}_{0,\infty}$ measurable. We then have

$$\begin{aligned}&D(P^\mu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}\|P^\nu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}})\\&\quad - D(P^\mu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}\|P^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}})\\&= E^\mu[\log(f_1)] - E^\mu[\log(f_2)]\\&= E^\nu[f_1\log(f_1)] - E^\nu[f_2\log(f_2)]\\&= \int_\Omega f_1(\omega)\log(f_1(\omega))dP^\nu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}(\omega)\\&\quad - \int_\Omega f_2(\omega)\log(f_2(\omega))dP^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}(\omega)\\&= \int_A f_1(\omega)\log(f_1(\omega))dP^\nu|_{\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{X}}_{n,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}}(\omega)\\&\quad - \int_A f_2(\omega)\log(f_2(\omega))dP^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}(\omega)\\&= \int_A f_1(\omega)\log(f_1(\omega))dP^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}(\omega)\\&\quad - \int_A f_2(\omega)\log(f_2(\omega))dP^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,\infty}}(\omega) = 0.\end{aligned}$$

Therefore, if the relative entropy of the filter is ever finite, then total variation merging in expectation is equivalent to merging in relative entropy.

V. Conclusion

We presented a notion of stochastic observability for nonlinear systems. This notion is explicit, is relatively easily computed due to its functional approximation formulation, and is shown via examples to be applicable to a large class of systems. The implications of this definition for filter stability were presented in detail. Further relations under various stability criteria and implications were studied.

Appendix
Supporting Results for Section IV-E

We present a number of supporting results. The approach for these build on similar arguments in [14] and [55]. The proofs here are kept brief due to space constraints or omitted.

Lemma A.1: Assume $\mu \ll \nu$. For any sigma field $\mathcal{G}\subseteq\mathcal{F}^{\mathcal{X}}_{0,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}$, we have

$$\frac{dP^\mu|_{\mathcal{G}}}{dP^\nu|_{\mathcal{G}}} = E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle|\mathcal{G}\right] \quad P^\mu \text{ a.s.}$$

Lemma A.2: Assume $\mu \ll \nu$. For any two sigma fields $\mathcal{G}_1, \mathcal{G}_2 \subset \mathcal{F}^{\mathcal{X}}_{0,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}$, let $P^\mu|_{\mathcal{G}_1}|\mathcal{G}_2$ represent the probability measure P^μ restricted to $\mathcal{G}_1$, conditioned on field $\mathcal{G}_2$. We then have

$$\frac{dP^\mu|_{\mathcal{G}_1}|\mathcal{G}_2}{dP^\nu|_{\mathcal{G}_1}|\mathcal{G}_2} = \frac{E^\nu[\frac{d\mu}{d\nu}(X_0)|\mathcal{G}_1\vee\mathcal{G}_2]}{E^\nu[\frac{d\mu}{d\nu}(X_0)|\mathcal{G}_2]} \quad P^\mu \text{ a.s.}$$

Lemma A.3: Assume $\mu \ll \nu$, for any two sigma fields $\mathcal{G}_1, \mathcal{G}_2 \subset \mathcal{F}^{\mathcal{X}}_{0,\infty}\vee\mathcal{F}^{\mathcal{Y}}_{0,\infty}$, we have P^μ a.s.

$$\begin{aligned}&\|P^\mu|_{\mathcal{G}_1}|\mathcal{G}_2 - P^\nu|_{\mathcal{G}_1}|\mathcal{G}_2\|_{TV}\\&= \frac{E^\nu\left[\left|E^\nu\left[\frac{d\mu}{d\nu}(X_0)|\mathcal{G}_1\vee\mathcal{G}_2\right] - E^\nu\left[\frac{d\mu}{d\nu}(X_0)|\mathcal{G}_2\right]\right|\middle|\mathcal{G}_2\right]}{E^\nu\left[\frac{d\mu}{d\nu}(X_0)|\mathcal{G}_2\right]}\end{aligned}$$

For the specific case of the nonlinear filter, that is, $\mathcal{G}_1 = \mathcal{F}^{\mathcal{X}}_n$ and $\mathcal{G}_2 = \mathcal{F}^{\mathcal{Y}}_{0,n}$, the results presented above imply the following known results in the literature.

Lemma A.4 (see [55, Lemma 5.6]): Assume $\mu \ll \nu$. Then, we have that $\pi_n^\mu \ll \pi_n^\nu$ a.s., and we have

$$\frac{d\pi_n^\mu}{d\pi_n^\nu}(x) = \frac{E^\nu[\frac{d\mu}{d\nu}(X_0)|Y_{[0,n]}, X_n = x]}{E^\nu[\frac{d\mu}{d\nu}(X_0)|Y_{[0,n]}]} \quad P^\mu \text{ a.s.} \tag{15}$$

Lemma A.5 (see [14, eq. 1.10]): The filter merges in total variation in expectation if and only if P^ν a.s.

$$E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle|\bigcap_{n\geq 0}\mathcal{F}^{\mathcal{Y}}_{0,\infty}\vee\mathcal{F}^{\mathcal{X}}_{n,\infty}\right] = E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle|F^{\mathcal{Y}}_{0,\infty}\right]. \tag{16}$$

Since our results apply to any general sigma field, not just the fields used in the analysis of the filter, we can study the predictor process to establish Lemmas A.6–A.8, in the following.

Lemma A.6: Assume $\mu \ll \nu$. Then, we have that $\pi_{n-}^\mu \ll \pi_{n-}^\mu$ P^μ a.s., and we have

$$\frac{d\pi_{n-}^\mu}{d\pi_{n-}^\nu}(x) = \frac{E^\nu[\frac{d\mu}{d\nu}(X_0)|Y_{[0,n-1]}, X_n = x]}{E^\nu[\frac{d\mu}{d\nu}(X_0)|Y_{[0,n-1]}]} \quad P^\mu \text{ a.s.} \tag{17}$$

Proof: These results become clear from Lemma A.2 when we state the predictor as P^μ restricted to $\mathcal{F}^{\mathcal{X}}_n$ conditioned on $\mathcal{F}^{\mathcal{Y}}_{0,n-1}$.

Lemma A.7: Assume $\mu \ll \gamma$ for some measure γ. We can express (18) shown at the top of the next page.

Proof: By Lemma A.3, we can write unnumbered equation shown at the top of the next page. Since Y_n is a function of X_n and the random noise Z_n, which is independent of X_n and past $Y_{[0,n-1]}$ measurements, we have that $\sigma(Y_{[0,n-1]},$

$$\|\pi^\mu_{n-} - \pi^\gamma_{n-}\|_{TV} = \frac{E^\gamma\left[\left|E^\gamma[\frac{d\mu}{d\gamma}(X_0)|Y_{[0,\infty)}, X_{[n,\infty)}] - E^\gamma[\frac{d\mu}{d\gamma}(X_0)|Y_{[0,n-1]}]\right|\middle| Y_{[0,n-1]}\right]}{E^\gamma\left[\frac{d\mu}{d\gamma}(X_0)\middle| Y_{[0,n-1]}\right]}. \tag{18}$$

$$\|\pi^\mu_{n-} - \pi^\gamma_{n-}\|_{TV} = \frac{E^\gamma\left[|E^\gamma[\frac{d\mu}{d\gamma}(X_0)|Y_{[0,n-1]}, X_n] - E^\gamma[\frac{d\mu}{d\gamma}(X_0)|Y_{[0,n-1]}]|\middle| Y_{[0,n-1]}\right]}{E^\gamma\left[\frac{d\mu}{d\gamma}(X_0)\middle| Y_{[0,n-1]}\right]}.$$

$X_n) = \sigma(Y_{[0,n]}, X_n)$. Further, by the Markov property, we have that we have that $(X_{[0,n-1]}, Y_{[0,n-1]})$ are independent of $(X_{[n+1,\infty)}, Y_{[n+1,\infty)})$ conditioned on (X_n, Y_n); therefore, we can state

$$E^\gamma\left[\frac{d\mu}{d\nu}(X_0)|Y_{[0,n-1]}, X_n\right] = E^\gamma\left[\frac{d\mu}{d\nu}(X_0)|Y_{[0,\infty)}, X_{[n,\infty)}\right].$$

Lemma A.8: The predictor merges in total variation in expectation if and only if P^ν a.s.

$$E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle|\bigcap_{n\geq 1}\mathcal{F}^{\mathcal{Y}}_{0,\infty}\vee\mathcal{F}^{\mathcal{X}}_{n,\infty}\right] = E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle| F^{\mathcal{Y}}_{0,\infty}\right]. \tag{19}$$

Proof: Building on the proof of Lemma A.7, we have

$$E^\mu\left[\|\pi^\mu_{n-} - \pi^\nu_{n-}\|_{TV}\right] = E^\nu\left[\frac{dP^\mu|_{\mathcal{F}^{\mathcal{Y}}_{0,n-1}}}{dP^\nu|_{\mathcal{F}^{\mathcal{Y}}_{0,n-1}}}\|\pi^\mu_{n-} - \pi^\nu_{n-}\|_{TV}\right]$$
$$= E^\nu\left[E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle| Y_{[0,n-1]}\right]\|\pi^\mu_{n-} - \pi^\nu_{n-}\|_{TV}\right]$$
$$= E^\nu\left[E^\nu\left[\left|E^\nu\left[\frac{d\mu}{d\nu}(X_0)\middle| Y_{[0,\infty)}, X_{[n,\infty)}\right]\right.\right.\right.$$
$$\left.\left. - E^\nu\left[\frac{d\mu}{d\nu}(X_0)|Y_{[0,n-1]}\right]||Y_{[0,n-1]}\right]\right]$$
$$= E^\nu\left[\left|E^\nu\left[\frac{d\mu}{d\nu}(X_0)|Y_{[0,\infty)}, X_{[n,\infty)}\right]\right.\right.$$
$$\left.\left. - E^\nu\left[\frac{d\mu}{d\nu}(X_0)|Y_{[0,n-1]}\right]\right|\right].$$

We then see that $A_n = E^\nu[\frac{d\mu}{d\nu}(X_0)|Y_{[0,n-1]}]$ is a nonnegative uniformly integrable martingale adapted to the increasing filtration $\mathcal{F}^{\mathcal{Y}}_{0,n-1}$. Hence, the limit as $n\to\infty$ in $L^1(P^\nu)$ is $E^\nu[\frac{d\mu}{d\nu}(X_0)|\mathcal{F}^{\mathcal{Y}}_{0,\infty}]$. Similarly, we can view $B_n = E^\nu[\frac{d\mu}{d\nu}(X_0)|Y_{[0,\infty)}, X_{[n,\infty)}]$ as backward nonnegative uniformly integrable martingale with respect to the decreasing sequence of filtrations $\mathcal{F}^{\mathcal{Y}}_{0,\infty}\vee\mathcal{F}^{\mathcal{X}}_{n,\infty}$. Then, by the backward martingale convergence theorem, the limit as $n\to\infty$ in $L^1(P^\nu)$ is $E^\nu[\frac{d\mu}{d\nu}(X_0)|\bigcap_{n=1}^\infty\mathcal{F}^{\mathcal{Y}}_{1,\infty}\vee\mathcal{F}^{\mathcal{X}}_{n,\infty}]$. It is then clear the total variation in expectation is zero if and only if (19) holds.

References

[1] B. D. O. Anderson and J. B. Moore, "Detectability and stabilizability of time-varying discrete-time linear systems," *SIAM J. Control Optim.*, vol. 19, no. 1, pp. 20–32, 1981.

[2] Y. Bar-Shalom and E. Tse, "Dual effect, certainty equivalence, and separation in stochastic control," *IEEE Trans. Autom. Control*, vol. AC-19, no. 5, pp. 494–500, Oct. 1974.

[3] A. R. Barron, "Information theory and martingales," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 1991, pp. 23–29. [Online]. Available: http://www.stat.yale.edu/~arb4/publications_files/information%20theory%20and%20martingales.pdf

[4] A. R. Barron, "Limits of information, Markov chains, and projections," in *Proc. IEEE Int. Symp. Inf. Theory*, 2000, p. 25.

[5] P. Billingsley, *Probability and Measure*, 2nd ed. New York, NY, USA: Wiley, 1986.

[6] P. Billingsley, *Convergence of Probability Measures*, 2nd ed. New York, NY, USA: Wiley, 1999.

[7] D. Blackwell and L. Dubins, "Merging of opinions with increasing information," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 882–886, 1962.

[8] V. I. Bogachev, *Measure Theory*. Berlin, Germany: Springer-Verlag, 2007.

[9] V. S. Borkar, "White-noise representations in stochastic realization theory," *SIAM J. Control Optim.*, vol. 31, pp. 1093–1102, 1993.

[10] P. E. Caines, *Linear Stochastic Systems*. New York, NY, USA: Wiley, 1988.

[11] C. T. Chen, *Linear Systems Theory and Design*. Oxford, U.K.: Oxford Univ. Press, 1999.

[12] P. Chigansky and R. van Handel, "A complete solution to Blackwell's unique ergodicity problem for hidden Markov chains," *Ann. Appl. Probability*, vol. 20, no. 6, pp. 2318–2345, 2010.

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

[14] P. Chigansky, R. Liptser, and R. van Handel, "Intrinsic methods in filter stability," in *Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovskii Eds., Oxford Univ. Press, 2009.

[15] J. M. C. Clark, D. L. Ocone, and C. Coumarbatch, "Relative entropy and error bounds for filtering of Markov processes," *Math. Control, Signals Syst.*, vol. 12, no. 4, pp. 346–360, 1999.

[16] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.

[17] A. D'Aristotile, P. Diaconis, and D. Freedman, "On merging of probabilities," *Sankhyā: Indian J. Statist., Ser. A*, vol. 50, pp. 363–380, 1988.

[18] R. M. Dudley, *Real Analysis and Probability*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[19] J. Dugundji, "An extension of Tietze's theorem," *Pacific J. Math.*, vol. 1, no. 3, pp. 353–367, 1951.

[20] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, Hoboken, NJ, USA: Wiley, 2009.

[21] E. A. Feinberg, P. O. Kasyanov, and M. Z. Zgurovsky, "Partially observable total-cost Markov decision process with weakly continuous transition probabilities," *Math. Operations Res.*, vol. 41, no. 2, pp. 656–681, 2016.

[22] J. Fritz, "An information-theoretical proof of limit theorems for reversible Markov processes," in *Proc. Trans. 6th Prague Conf. Inf. Theory, Statist. Decis. Functions, Random Processes*, 1973, pp. 183–197.

[23] I. I. Gihman and A. V. Skorohod, *Controlled Stochastic Processes*. Berlin, Germany: Springer, 2012.

[24] N. Golowich, A. Moitra, and D. Rohatgi, "Planning in observable POMDPs in quasipolynomial time," 2022, *arXiv:2201.04735*.

[25] R. M. Gray, *Entropy and Information Theory*. Berlin, Germany: Springer, 2011.

[26] P. Harremoës and K. K. Holst, "Convergence of Markov chains in information divergence," *J. Theor. Probability*, vol. 22, pp. 186–202, 2009.

[27] P. Harremoës, O. Johnson, and I. Kontoyiannis, "Thinning, entropy, and the law of thin numbers," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4228–4244, Sep. 2010.

[28] R. Hermann and A. Krener, "Nonlinear controllability and observability," *IEEE Trans. Autom. Control*, vol. AC-22, no. 5, pp. 728–740, Oct. 1977.

[29] C. C. White III and W. T. Scherer, "Finite-memory suboptimal design for partially observed Markov decision processes," *Operations Res.*, vol. 42, no. 3, pp. 439–455, 1994.

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

[31] A. D. Kara, N. Saldi, and S. Yüksel, "Weak Feller property of non-linear filters," *Syst. Control Lett.*, vol. 134, pp. 104–512, 2019.

[32] A. D. Kara and S. Yüksel, "Robustness to incorrect priors in partially observed stochastic control," *SIAM J. Control Optim.*, vol. 57, no. 3, pp. 1929–1964, 2019.

[33] A. D. Kara and S. Yüksel, "Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability," *Math. Operations Res.*, 2022, *arXiv:2103.12158*.

[34] A. D. Kara and S. Yüksel, "Near optimality of finite memory feedback policies in partially observed Markov decision processes," *J. Mach. Learn. Res.*, vol. 23, no. 11, pp. 1–46, 2022.

[35] J.-W. Kim and P. G. Mehta, "Duality for nonlinear filtering I: Observability," *IEEE Trans. Automat. Control*, 2023.

[36] J.-W. Kim, P. G. Mehta, and S. P. Meyn, "What is the Lagrangian for nonlinear filtering," in *Proc. IEEE 58th Conf. Decis. Control*, 2019, pp. 1607–1614.

[37] H. Kunita, "Asymptotic behavior of the nonlinear filtering errors of Markov processes," *J. Multivariate Anal.*, vol. 1, no. 4, pp. 365–393, 1971.

[38] H. J. Kushner, "A partial history of the early development of continuous-time nonlinear stochastic systems theory," *Automatica*, vol. 50, no. 2, pp. 303–334, 2014.

[39] H. J. Kushner, *Introduction to Stochastic Control Theory*. New York, NY, USA: Holt, Rinehart and Winston, 1972.

[40] T. Linder and S. Yüksel, "On optimal zero-delay coding of vector Markov sources," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 5975–5991, Oct. 2014.

[41] A. R. Liu, "Stochastic observability, reconstructibility, controllability, and reachability," Ph.D. Thesis, Dept. Mechanical Eng., UC San Diego, La Jolla, CA, USA, 2011.

[42] A. R. Liu and R. R. Bitmead, "Observability and reconstructibility of hidden Markov models: Implications for control and network congestion control," in *Proc. IEEE 49th Conf. Decis. Control*, 2010, pp. 918–923.

[43] C. McDonald and S. Yüksel, "Exponential filter stability via Dobrushin's coefficient," *Electron. Commun. Probability*, vol. 25, pp. 1–13, 2020.

[44] C. McDonald and S. Yüksel, "Robustness to incorrect priors and controlled filter stability in partially observed stochastic control," *SIAM J. Control Optim.*, vol. 60, no. 2, pp. 842–870, 2022.

[45] S. P. Meyn and R. Tweedie, *Markov Chains and Stochastic Stability*. Berlin, Germany: Springer-Verlag, 1993.

[46] H. Nijmeijer, "Observability of autonomous discrete time non-linear systems: A geometric approach," *Int. J. Control*, vol. 36, no. 5, pp. 867–874, 1982.

[47] D. L. Ocone, "Asymptotic stability of Beneš filters," *Stochastic Anal. Appl.*, vol. 17, no. 6, pp. 1053–1074, 1999.

[48] A. S. Reddy and A. Apte, "Stability of non-linear filter for deterministic dynamics," *Foundations Data Sci.*, vol. 3, no. 3, pp. 647–675, 2021.

[49] W. Rudin, *Real and Complex Analysis*. New York, NY, USA: Tata McGraw-Hill Education, 2006.

[50] E. D. Sontag, "A concept of local observability," *Syst. Control Lett.*, vol. 5, no. 1, pp. 41–47, 1984.

[51] R. B. Sowers and A. M. Makowski, "Discrete-time filtering for linear systems with non-Gaussian initial conditions: Asymptotic behavior of the difference between the MMSE and LMSE estimates," *IEEE Trans. Autom. Control*, vol. 37, no. 1, pp. 114–120, Jan. 1992.

[52] V. A. Ugrinovskii, "Observability of linear stochastic uncertain systems," *IEEE Trans. Autom. Control*, vol. 48, no. 12, pp. 2264–2269, Dec. 2003.

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

[54] R. van Handel, "Observability and nonlinear filtering," *Probability Theory Related Fields*, vol. 145, pp. 35–74, 2009.

[55] R. van Handel, "The stability of conditional Markov processes and Markov chains in random environments," *Ann. Probability*, vol. 37, no. 5, pp. 1876–1925, 2009.

[56] R. van Handel, "Uniform observability of hidden Markov models and filter stability for unstable signals," *Ann. Appl. Probability*, vol. 19, no. 3, pp. 1172–1199, 2009.

[57] R. van Handel, "Nonlinear filtering and systems theory," in *Proc. 19th Int. Symp. Math. Theory Netw. Syst.*, 2010.

Curtis McDonald received the B.A.Sc. degree in mathematics and engineering and the master's degree in mathematics and statistics from Queen's University, Kingston, ON, Canada. He is currently working toward the Ph.D. degree with the Statistics and Data Science Department, Yale University, New Haven, CT, USA. His thesis research focuses on Markov Chain Monte Carlo techniques and Bayesian methods.

His research interests include mathematical statistics, stochastic control theory, information theory, and probability.

Serdar Yüksel (Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey in 2001, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois Urbana-Champaign, Champaign, IL, USA, in 2003 and 2006, respectively.

He was a Postdoctoral Researcher with Yale University, New Haven, CT, USA, before joining Queen's University, Kingston, ON, Canada, as an Assistant Professor with the Department of Mathematics and Statistics, where he is currently a Professor. His research interests include stochastic control, decentralized control, information theory, and probability.