Change Detection in Partially Observed Nonlinear Dynamic Systems with Unknown Change Parameters

Namrata Vaswani**
Dept. of Electrical & Computer Engineering and Center for Automation Research,
University of Maryland, College Park, MD 20742, USA
namrata@cfar.umd.edu

***Abstract*— We study the change detection problem in partially observed nonlinear dynamic systems. We assume that the change parameters are unknown and the change could be gradual (slow) or sudden (drastic). For most nonlinear systems, no finite dimensional filters exist and approximation filtering methods like the Particle Filter are used. Even when change parameters are unknown, drastic changes can be detected easily using the increase in tracking (output) error or the negative log of observation likelihood (OL). But slow changes usually get missed. We propose in this paper, a statistic for slow change detection which turns out to be the same as the Kerridge Inaccuracy between the posterior state distribution and the normal system prior. We show asymptotic convergence (under certain assumptions) of the bounding, modeling and particle filtering errors in its approximation using a particle filter optimal for the normal system. We also demonstrate using the bounds on the errors that our statistic works in situations where observation likelihood (OL) fails and vice versa.**

I. Introduction

Change or abnormality detection is required in many practical problems arising in quality control, flight control, fault detection and in surveillance problems like abnormal activity detection [1]. In most cases, the underlying system in its normal state can be modeled as a parametric stochastic model (which may be linear or nonlinear). The observations are usually noisy (making the system partially observed) and the transformation between the observation and the state may also be linear or nonlinear. Such a system, in the most general case, forms a Partially Observed Non-Linear Dynamical (PONLD) system and in general can be tracked/filtered (approximately) using a finite dimensional Particle Filter (PF) [2]. We study here the change detection problem in PONLD systems when change parameters are unknown and the change could be slow or drastic.

If the change is drastic, the likelihood of observations under the normal (unchanged) model will reduce (OL which is its negative log will increase) or equivalently the particle filter, which is optimal for the normal system, will lose track. Thus OL can be used to detect this change. But due to asymptotic stability [3], the particle filter is able to track slow changes and hence these get missed by OL. We propose, in this work, a statistic for slow change detection, called ELL, which in fact can be estimated correctly for the changed system (using a particle filter optimal for the normal system) only because of asymptotic stability.

ELL or Expected (negative) Log Likelihood at time t, is the *expectation w.r.t. the posterior distribution, of the negative log of the prior likelihood of the state, under the no change hypothesis (H_0)*. In [4], the Kerridge Inaccuracy [5] between the empirical distribution of a set of N i.i.d. observations and their actual pdf is shown to be the same as the average negative log-likelihood. We show here the equivalence between ELL and Kerridge Inaccuracy between the posterior and prior state distributions. We study the errors in ELL approximation (bounding error, model error and PF error) and show their asymptotic convergence to zero (as the bound, time and number of particles go to infinity). The error upper bounds are then used to show complementary behavior of ELL and OL for slow and drastic changes. Thus *for changes where the rate of change could be anywhere from slow to drastic, we propose to use a combination of ELL and OL.*

A. Related Work

Online detection of changes for partially observed *linear* dynamical systems has been studied extensively. For *known changed system parameters*, the CUSUM [6] algorithm can be used directly. For *unknown changed system parameters*, the Generalized Likelihood Ratio Test can be used whose solution for *linear* systems in well known [6]. When a nonlinear system experiences a change, linearization techniques like Extended Kalman Filtering and change detection methods for linear systems are the main tools [6]. Linearization techniques are computationally efficient but are not always applicable (require a good initial guess at each time step and hence are not robust to noise spikes).

[7] is an attempt to use a Particle Filtering (PF) approach for *sudden* change detection in Partially Observed *Non-Linear* Dynamical (PONLD) systems without linearization. It assumes that *the parameters of the changed system are known* and defines a modification of the CUSUM change detection statistic that can be efficiently evaluated using particle filters. Both CUSUM and [7] are based on the current observation's likelihood ratio, given past observations. Tracking error (or output error) [8] which is the distance (usually Euclidean distance) between the current observation and its prediction based on past observations can also be used for *sudden* change detection and it *does not require knowledge of the changed system parameters*. An entirely different class of approaches (e.g. see [9]) used extensively with particle filters uses a *discrete state variable* to denote the mode that the system is operating in. But this approach also assumes *known* change parameters. In this case a change is detected by looking the expected or most probable value of the state variable.

There has been a lot of recent research on stability of the optimal nonlinear filter. Asymptotic stability results w.r.t. initial condition were first proved in [10]. The Hilbert projective metric has been used to prove stability w.r.t. the initial condition and also w.r.t. the model [11], [12]. New approaches have been proposed recently for noncompact state spaces [13], [14]. The results for stability w.r.t. the model have been used to prove convergence of the particle filtering estimate of the posterior with number of particles, $N \to \infty$ [3], [15]. We use in this paper results from [3] in which the authors have replaced the mixing transition kernel assumption required for proving stability with a much weaker mixing unnormalized filter kernel assumption.

**The author would like to acknowledge Prof. Rama Chellappa and Prof. Adrian Papamarcou for their valuable suggestions.

B. The PONLD Model

We assume that we have a $\Re^{n_x}$ valued state process $X = \{X_t\}$ and an $\Re^{n_y}$ valued observation process $Y = \{Y_t\}$[1]. The system (or state) process $\{X_t\}$ for the original system is assumed to be a Markov process with state transition kernel $Q_t(x_t, dx_{t+1})$ and the observation process is defined by $Y_t = h_t(X_t) + w_t$ where w_t is an i.i.d. noise process and h_t is, in general, a nonlinear function. The prior initial state distribution, denoted by $\pi_0(dx)$, the conditional distribution of observation given state, $G_t(x_t, dy_t)$, with pdf given by $g_t(x, Y_t) \triangleq \psi_t(x)$, and the state transition kernel, $Q_t(x_t, dx_{t+1})$, are known and assumed absolutely continuous[2]. A **non-linear filter** estimates the posterior probability distribution of the state at time t given the observations up to time t, $Pr(X_t \in dx_t|Y_{1:t}) \triangleq \pi_t(dx_t)$. We assume that the normal (original/unchanged) system has state transition kernel Q_t^0. A change in the system model begins to occur at some time t_c and lasts till a final time t_f. In the time interval, $[t_c, t_f]$, the state transition kernel is Q_t^c and after t_f it again becomes Q_t^0. Both Q_t^c and the change start and end times t_c, t_f are assumed unknown. The aim is to detect the change, with minimum delay.

The paper is organized as follows: ELL, its relation with Kerridge Inaccuracy and the motivation for using it for gradual change detection is discussed in Section III. In Section IV, we study the errors in approximating the ELL and state our asymptotic convergence theorems. In Section V, we analyze the implications of our results from Section IV for finite time and finite number of particles and discuss situations where ELL would detect changes better than OL and vice versa. We present simulation results and results on a real abnormal activity detection problem in Section VI and give conclusions and future work in Section VII.

II. Preliminaries

We briefly discuss below some notation and definitions of terms used in the rest of the paper. We then explain in Section II-B, the optimal nonlinear filter and its approximation using a particle filter.

A. Notation and Definitions

We use H_0 to denote the original or unchanged system hypothesis and H_c to denote the changed system hypothesis. Also, the superscript ^{c} is used to denote any parameter related to the changed system, ^{0} for the original system and ^{c,0} for the case when the observations of the changed system are filtered using a filter designed for the original system[3]. Thus the posteriors, $\pi_t^{0,0}(dx) = Pr(X_t \in dx|Y_{1:t}^0, H_0)$ (also denoted by π_t^0), $\pi_t^{c,c}(dx) = Pr(X_t \in dx|Y_{1:t}^c, H_c)$ (also denoted by π_t^c) and $\pi_t^{c,0}(dx) = Pr(X_t \in dx|Y_{1:t}^c, H_0)$ where

$$\begin{aligned} Y_{1:t}^c &= (Y_{1:t_c-1}^0, Y_{t_c:t}^c), \ \forall t \le t_f \\ &= (Y_{1:t_c-1}^0, Y_{t_c:t_f}^c, Y_{t_f+1:t}^0), \ \forall t > t_f. \end{aligned} \tag{1}$$

The prior state distribution at time t, $(Q_t^0, ...Q_1^0\pi_0)(dx)$ has pdf $p_t(x)$ while the changed system's prior state distribution, $(Q_t^0, ...Q_{t_f}^c, ..Q_{t_c}^c...Q_1^0\pi_0)(dx)$ has pdf $p_t^c(x)$. In a lot of cases (for example if the system model is linear Gaussian with Gaussian initial state pdf) it is possible to define the pdfs $p_t(x)$ and $p_t^c(x)$ in closed form. In cases where it cannot be defined closed form, it can be approximated by a single or a mixture of Gaussians (depending on whether it is unimodal or multimodal).

Note that throughout the paper, **"event occurs $a.s.$"** refers to the event occurring almost surely w.r.t. the measure corresponding to the probability distribution of $Y_{1:t}$. Also, E_μ denotes expectation under the measure μ, for example E_{π_t} is expectation under the posterior state distribution. E_Y denotes expectation under the distribution of the random variable Y, for example $E_{Y_{1:t}}$ denotes expectation under the distribution of the observation sequences. Finally, Ξ_{pf} denotes averaging over different realizations of the particle filter each of which produces a different random measure π_t^N[4].

With any nonnegative kernel, J, defined on the state space, E, is associated a nonnegative linear operator denoted by J and defined by $J\mu(dx') \triangleq \int_E \mu(dx)J(x, dx')$ for any nonnegative measure μ. Also, $(.,.)$ is the inner product notation.

Definition 1: The **unnormalized kernel describing the optimal filter** for a system with state transition kernel Q_t and probability of observation given state ψ_t, is given by $R_t(x, dx') = Q_t(x, dx')\psi_t(x')$. So $R_t^0 = Q_t^0\psi_t^0$ is the unnormalized optimal filter for original system observations, $R_t^c = Q_t^c\psi_t^c$ is the unnormalized optimal filter for the changed system observations while $R_t^{c,0} = Q_t^0\psi_t^c$ is the unnormalized filter (not optimal) for the changed system observations using original system transition kernel (this is what is done in practice since Q_t^c is assumed unknown).

Definition 2: A nonnegative kernel J defined on E is **mixing** if there exists a constant, $0 < \epsilon \le 1$ and a nonnegative measure λ s.t. $\epsilon\lambda(A) \le J(x, A) \le \frac{1}{\epsilon}\lambda(A) \ \forall x \in E$ and for any Borel subset $A \subset E$. A (time) sequence of mixing kernels $\{J_t\}$ is said to be **uniformly mixing** if $\epsilon = \sup_t \epsilon_t > 0$.

Definition 3: The **Birkhoff's contraction coefficient** of any kernel J is, $\tau(J) = \sup_{0 \le h(\mu,\mu') < \infty} \frac{h(J\mu, J\mu')}{h(\mu,\mu')} = tanh[\frac{1}{4}\sup_{\mu,\mu'} h(J\mu, J\mu')]$. h here denotes the Hilbert metric which is defined and explained in [3]. $\tau(J) \le 1$ always and if J is mixing, $\tau(J) \le \tilde{\tau}(J) < 1$ where $\tilde{\tau}(J) \triangleq \frac{1-\epsilon^2}{1+\epsilon^2} < 1$. We denote $\tau(R_t)$ by τ_t and $\epsilon(R_t)$ by ϵ_t. Note that R_t depends on Y_t and hence τ_t and ϵ_t are, in general, random variables. So a correct statement would be that R_t is a.s. mixing ($\epsilon_t > 0, a.s.$ and $\tau_t < 1, a.s.$).

B. Non-linear Filtering

The problem of nonlinear filtering is to compute at each time t, the conditional probability distribution, of the state X_t given the observation sequence $Y_{1:t} = (Y_1, Y_2, ...Y_t)$, $\pi_t(dx) = Pr(X_t \in dx|Y_{1:t})$. The transition from π_{t-1} to π_t is defined using the Bayes recursion as follows:

$$\pi_{t-1} \text{ ---> } \pi_{t|t-1} = Q_t\pi_{t-1} \text{ ---> } \pi_t = \frac{\psi_t\pi_{t|t-1}}{(\pi_{t|t-1}, \psi_t)}$$

Now if the system and observation models are linear Gaussian, the posteriors would also be Gaussian and can be evaluated in closed form (Kalman filter). For nonlinear or nonGaussian system or observation model, except in very special cases, the filter is infinite dimensional. The Particle Filter [15] is a sequential monte

[1]We use the subscript 't' (e.g. X_t, Y_t) instead of 'n' for (discrete) time instants, to avoid confusion with N used for number of particles in Particle Filtering

[2]Note that for ease of notation, we denote the pdf either by the same symbol or by the lowercase of the probability distribution symbol

[3]Even if ^{0} is omitted, but there is no ^{c}, it denotes the original system.

[4]expectation under the probability distribution of the random measure π_t^N or equivalently of the random particles, $\{x_t^{(i)}\}_{i=1}^N$.

carlo method for nonlinear filtering which was first introduced in [2] as Bayesian Bootstrap Filtering.

Particle Filtering: A particle filter (PF) [15] is a recursive algorithm which produces at each time t, a cloud of N particles $\{x_t^{(i)}\}$ whose empirical measure, π_t^N (which is a random measure), closely "follows" π_t. It starts with sampling N times from π_0 to approximate it by $\pi_0^N(dx) \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{x_0^{(i)}}(dx)$. The Bayes recursion then runs as follows:

$$\pi_{t-1}^N \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{x_{t-1}^{(i)}}(dx) \longrightarrow \pi_{t|t-1}^N \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{\bar{x}_t^{(i)}}(dx)$$

$$\longrightarrow \bar{\pi}_t^N \triangleq \frac{1}{N}\sum_{i=1}^N w_t^{(i)}\delta_{\bar{x}_t^{(i)}}(dx) \longrightarrow \pi_t^N \triangleq \sum_{i=1}^N \delta_{x_t^{(i)}}(dx)$$

$$\begin{aligned} \text{where} \quad & \bar{x}_t^{(i)} \sim Q_t(x_{t-1}^{(i)}, dx), \\ & x_t^{(i)} \sim \text{Multinomial}(\{\bar{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^N) \\ & w_t^{(i)} \triangleq \frac{\psi_t(\bar{x}_t^{(i)})}{(\pi_{t|t-1}^N, \psi_t(\bar{x}_t^{(i)}))} \end{aligned} \tag{2}$$

Note that the last step is aimed at reducing degeneracy of the particles. The samples $\bar{x}_t^{(i)}$ are resampled assuming a multinomial distribution proportional to their weights, $w_t^{(i)}$, so that particles with very low weights get eliminated while those with higher weights get repeated in proportion to their weights.

III. The ELL Statistic

"Expected (negative) Log Likelihood" or ELL at time t, is the *expectation w.r.t. the posterior distribution (π_t), of the negative log of the prior likelihood of the state, under the no change hypothesis (H_0)*, i.e.

$$ELL(Y_{1:t}) \triangleq E_{\pi_t}[-\log p_t^0(x)]. \tag{3}$$

For systems where exact filters do not exist and a PF is used to estimate π_t, the estimate of ELL using the empirical distribution π_t^N becomes $ELL^N = \frac{1}{N}\sum_{i=1}^N[-\log p_t^0(x^{(i)})]$.

It is interesting to note that ELL as defined above is also the Kerridge Inaccuracy [5] between the posterior and prior state pdf. The **Kerridge inaccuracy** (KI) between two pdfs p, q, i.e. $K(p,q) = \int p(x)[-\log q(x)]dx$ is a measure of inaccuracy between distributions (used in statistics) and was first defined by Kerridge in [5]. We have $ELL(Y_{1:t}) \triangleq E_{\pi_t}[-\log p_t^0(x)] = K(\pi_t : p_t^0)$[5]. Henceforth, we denote $ELL(Y_{1:t}^0) = K(\pi_t^0 : p_t) \triangleq K_t^0$ and $ELL(Y_{1:t}^c) = K(\pi_t^c : p_t) \triangleq K_t^c$.

A. Motivation for ELL

The use of ELL (or equivalently KI) for partially observed systems is motivated by the use of log likelihood for hypothesis testing in the fully observed case. For a fully observed system (assuming h_t invertible), one could evaluate $X_t = h_t^{-1}(Y_t)$ from the observation Y_t and then $\log p_t(X_t)$ would be the log likelihood of state taking value X_t under H_0 (proportional to likelihood of Y_t under H_0). Thus if $Y_t = Y_t^0$, then its likelihood, and so also the likelihood of the state X_t, under H_0 will be larger than if $Y_t = Y_t^c$[6]. But for partially observed systems, X_t is not deterministic given $Y_{1:t}$. It is a random variable with distribution π_t. Hence we propose to replace log likelihood of the state by its expectation under π_t which is the ELL.

B. Why ELL (KI) works?

Taking expectation of $ELL(Y_{1:t}^0) = K(\pi_t^{0,0} : p_t^0)$ over normal observation sequences, we get

$$\begin{aligned} E_{Y_{1:t}^0}[ELL(Y_{1:t}^0)] &= E_{Y_{1:t}^0} E_{\pi_t^0}[-\log p_t^0(x)] \\ &= E_{p_t^0}[-\log p_t^0(x)] = K(p_t^0 : p_t^0) \triangleq EK_t^0 \end{aligned}$$

Similarly, for changed system observations, $E_{Y_{1:t}^c}[ELL(Y_{1:t}^c)] = K(p_t^c : p_t^0) \triangleq EK_t^c$, i.e. the expectation of ELL of changed system observations is actually the KI between the changed system prior, p_t^c, and original system prior, p_t^0, which will be larger than KI between p_t^0 and p_t^0 [4]. EK_t^c can be used as a measure of the change magnitude at time t (and dividing by the change duration until t gives a measure for the rate of change).

Now, ELL (KI) will detect the change when EK_t^c is "significantly" larger than EK_t^0. Setting the change threshold to $\kappa_t \triangleq EK_t^0 + 3\sqrt{VK_t^0}$, where $VK_t^0 = Var_{Y_{1:t}}(K_t^0)$, will ensure a false alarm probability less than 0.11 (0.05 if unimodal)[7]. By the same logic, if K_t^c is such that $EK_t^c - 3\sqrt{VK_t^c} > \kappa_t$ then the miss probability will also be less than 0.11 (0.05 if unimodal). Now evaluating VK_t^0 or VK_t^c analytically is not possible without having an analytical expression for π_t^0 or π_t^c. But we can bound VK_t^0 (and similarly VK_t^c) as follows (apply Jensen's inequality on z^2, which is a convex function, with $z = [-\log p_t(x)]$):

$$\begin{aligned} {K_t^0}^2 &= (E_{\pi_t}[-\log p_t(x)])^2 \le E_{\pi_t}[[-\log p_t(x)]^2] \\ \text{So,} \quad VK_t^0 &= Var_{Y_{1:t}^0}(K_t^0) = E_{Y_{1:t}}[{K_t^0}^2] - (EK_t^0)^2 \\ &\le E_{Y_{1:t}}[E_{\pi_t}[[-\log p_t(x)]^2]] - (EK_t^0)^2 \\ &= E_{p_t^0}[[-\log p_t^0(x)]^2] - (EK_t^0)^2 \end{aligned} \tag{4}$$

Example 1: Consider as an example the case where Q_t^0, Q_t^c and π_0 are linear Gaussian, so that p_t^0 and p_t^c are also Gaussian. Assume scalar state and observation and let the pdf of $Q_t(x, dx')$ is $\mathcal{N}(x, \sigma_{noise}^2)$ and pdf of $Q_t^c(x, x')$ is $\mathcal{N}(x + \Delta a, \sigma_{noise}^2)$. Thus p_t^0 is $\mathcal{N}(0, \sigma_t^2)$ and p_t^c is $\mathcal{N}(a_t, \sigma_t^2)$ where $a_t = t\Delta a, \sigma_t^2 = t\sigma_{noise}^2$. The non-linearity (if any) is in the mapping from state to observation space. Then, it is easy to see that

$$\begin{aligned} EK_t^0 &= K(p_t^0 : p_t^0) = 0.5\log 2\pi\sigma_t^2 + 0.5 \\ EK_t^c &= K(p_t^c : p_t^0) = 0.5\log 2\pi\sigma_t^2 + 0.5\frac{{\sigma_t^c}^2 + a_t^2}{\sigma_t^2} \\ &= 0.5\log 2\pi\sigma_t^2 + 0.5 + 0.5\frac{a_t^2}{\sigma_t^2}, \quad \text{since } {\sigma_t^c}^2 = \sigma_t^2 \end{aligned} \tag{5}$$

$$\begin{aligned} VK_t^0 &\le E_{p_t^0}[[-\log p_t^0(x)]^2] - (EK_t^0)^2 = 0.5 \\ VK_t^c &\le E_{p_t^c}[[-\log p_t^0(x)]^2] - (EK_t^c)^2 = 0.5 + \frac{a_t^2}{\sigma_t^2} \end{aligned} \tag{6}$$

Thus the above analysis shows that the mean distance of K_t^c from threshold is

$$\gamma_t \triangleq EK_t^c - \kappa_t \ge 0.5\frac{a_t^2}{\sigma_t^2} - 3\sqrt{0.5} \tag{7}$$

Now the miss probability at time t will be less than 0.11 (0.05 if unimodal) if $\gamma_t > 3\sqrt{VK_t^c}$ which in this case simplifies to

[5]it is actually $K(\frac{d\pi_t}{dx} : p_t^0)$ but as mentioned earlier, we denote the density $\frac{d\pi_t}{dx}$ by the same symbol as the distribution

[6]In this case observation likelihood and state likelihood (ELL) are proportional.

[7]0.11 follows by Chebyshev inequality [16]. But if the pdf of $K_t^0(Y_{1:t})$ is unimodal, Gauss's inequality [16] can be applied to show that the probability is less than 0.05

$0.5r^2 - 3\sqrt{0.5} > 3r$ with $r = a_t/\sigma_t$. This of course is obtained using very loose bounds (loose variance bound and the loose Chebyshev or Gauss's inequality bound) and in practice changes get detected much faster.

IV. Errors in ELL Approximation

Now the above analysis assumes there are no errors in estimating K_t^0 and K_t^c which is true only if exact finite dimensional filters exist for a problem and correct models for the transition kernel and conditional probability of observation given state are used. For example the estimation of K_t^0 in the linear Gaussian case (Kalman filter). But in all other cases there are three kinds of errors: When we are trying to estimate K_t^c using the transition kernel for the original system, what we really evaluate is $K_t^{c,0} \triangleq E_{\pi_t^{c,0}}[-\log p_t^0(x)]$ instead of K_t^c (**model error**). We can use the asymptotic stability result from [3] to show (under certain assumptions) that this error goes to zero for large time instants, for posterior expectations of bounded functions of the state. But $K_t^{c,0} = E_{\pi_t^{c,0}}[-\log p_t^0(x)]$ and $[-\log p_t^0(x)]$ is an unbounded function. Considering its bounded approximation introduces **bounding errors** which go to zero as the bound goes to infinity. Also, when we use a particle filter with finite number of particles to approximate the optimal filter, **PF approximation error** is introduced. This error goes to zero as the number of particles goes to infinity.

Now, we quantify our claims. Our aim is to *either* show a result of the type $\lim_{M\to\infty}(\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^0 : p_t) - K(\pi_t^{0,N} : p_t^M)|]) = 0$ and $\lim_{M\to\infty}(\lim_{t\to\infty}(\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t^M)|])) = 0$, $a.s.$, where $p_t^M(x) \triangleq \max\{p_t(x), e^{-M}\}$[8]. *Or* show that under certain assumptions, $[-\log p_t(x)]$ is uniformly bounded for all t so that the outermost convergence with M follows trivially. We use the following two theorems from [3]:

Theorem 1: (Model error bound, Theorem 4.6 of [3])- If for all k, the kernel R_k is a.s. mixing ($\implies \epsilon_k > 0, a.s.$ & Birkhoff's contraction coefficient $\tau_k \le \tilde{\tau}_k(\epsilon_k) < 1, a.s.$), then the weak norm between the correct optimal filter density μ_t and the incorrect one μ_t' is upper bounded as follows:

$$\sup_{\phi:||\phi||_\infty \le 1} |(\mu_t - \mu_t', \phi)| \le \delta_t + \frac{\delta_{t-1}}{\epsilon_t^2} + \sum_{k=1}^{t-2} \tilde{\tau}_{t:k+3} \frac{\delta_k}{\epsilon_{k+1}^2 \epsilon_{k+2}^2} \triangleq \theta_t(\delta_k, \epsilon_k, 0 \le k \le n), a.s. \quad (8)$$

$$\text{where} \quad \delta_k \triangleq \sup_{\phi:||\phi||_\infty \le 1} |(\mu_k' - \bar{R}_k \mu_{k-1}', \phi)| \le 2 \quad (9)$$

Theorem 2: (PF error bound, Theorem 5.7 of [3])- If for all k, the kernel R_k is a.s. mixing ($\epsilon_k > 0, a.s.$ & $\tau_k \le \tilde{\tau}_k(\epsilon_k) < 1, a.s.$), and $\sup_{x \in E_{x,y}} \psi_k(x) < \infty, a.s.$, then the weak norm between the correct optimal filter density μ_t and the approximation μ_t^N (evaluated using the PF) is upper bounded as follows:

$$\begin{aligned}
&\sup_{\phi:||\phi||_\infty \le 1} \Xi_{pf}[|(\mu_t - \mu_t^N, \phi)|] \\
&\le \frac{2(\rho_t + \frac{\rho_{t-1}}{\epsilon_t^2} + \sum_{k=1}^{t-2} \tilde{\tau}_{t:k+3} \frac{\rho_k}{\epsilon_{k+1}^2 \epsilon_{k+2}^2})}{\sqrt{N}} \\
&\triangleq \frac{\beta_t(\rho_k, \epsilon_k, 0 \le k \le n)}{\sqrt{N}}, a.s. \quad (10)
\end{aligned}$$

$$\text{where} \quad \rho_k \triangleq \frac{\sup_{x\in E} \psi_k(x)}{\inf_{\mu \in \mathcal{P}(E)} (Q_k \mu, \psi_k)} < \infty, a.s. \quad (11)$$

Now we can claim the following three results under progressively weaker assumptions (see author's website for proofs, http://www.cfar.umd.edu/∼namrata/accproof.pdf, which have been omitted here due to lack of space)

Theorem 3: Assuming (i) Change occurs for only a finite time period $[t_c : t_f]$ and starting time $t_c \le T^* < \infty$; (ii) R_k^c is uniformly mixing for all k (with mixing parameters $\epsilon_k^c > \epsilon^c > 0$), R_k^0 is uniformly mixing for all k (with $\epsilon_k^0 > \epsilon^0 > 0$) and $R_k^{c,0} \triangleq Q_k^0(x, dx')\psi_k^c(x')$ is uniformly mixing for all k (with $\epsilon_k^{c,0} > \epsilon^{c,0} > 0$); (iii) $\sup_{x\in E_{x,y}} \psi_k(x) < \infty, a.s., \ \forall k$ [9] and (iv)(a) The posterior state space is uniformly compact for all t, i.e. $E_{x,y} \triangleq \{x \in E : \{\psi(y_t^0|x) > 0\} \ or \{\psi(y_t^c|x) > 0\} \ \text{ for some } t\}$ is a compact set, and (b) there exists $\alpha > 0$, s.t. $p_t(x) > \alpha, \ \forall x \in E_{x,y}, \ \forall t$; then the following result holds:

$$\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^0 : p_t) - K(\pi_t^{0,N} : p_t)|] = 0, a.s.$$
$$\lim_{t\to\infty}(\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t)|]) = 0, a.s.$$

Now assumption (iv) in the theorem above ensures that $[-\log p_t(x)]$ is uniformly bounded $\forall t$, so that theorems 1 and 2 can be applied to prove the result. But one can relax this assumption by defining a sequence of functions $\{[-\log p_t^M(x)]\}$ with $p_t^M(x) = \max\{p_t(x), e^{-M}\}$, s.t. $\lim_{M\to\infty}[-\log p_t^M(x)] = [-\log p_t(x)]$. Then by a simple extension of Monotone Convergence Theorem ([17], page 87) to functions which could be negative but are bounded from below, we have $\lim_{M\to\infty} K(\pi_t^c : p_t^M) = K(\pi_t^c : p_t)$. We then have the following result.

Theorem 4: Assuming (i), (ii), (iii) as in Theorem 3, and (iv) being replaced by the weaker assumption (iv)′: Convergence of $K(\pi_t^c : p_t^M)$ to $K(\pi_t^c : p_t)$ is uniform in t, we have

$$\lim_{M\to\infty}(\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^0 : p_t) - K(\pi_t^{0,N} : p_t^M)|]) = 0, a.s.$$
$$\lim_{M\to\infty}(\lim_{t\to\infty} \Xi_{pf}[(\lim_{N\to\infty} |K(\pi_t^c : p_t) - K(\pi_t^{c,0,N} : p_t^M)|])) = 0, a.s.$$

Theorem 5: If neither of (iv) or (iv)′ is assumed and (ii) is replaced by the weaker assumption (ii)′: The kernels $R_t^c, R_t^{c,0}, R_t^0$ are mixing (not uniformly mixing), then for normal observations, Theorem 4 still holds but for changed observations we are currently able to claim only the following finite time result: Given any $\Delta > 0$, there exists an $M_{t,\Delta}$ s.t.

$$\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^{c,0,N} : p_t^{M_{t,\Delta}}) - K(\pi_t^c : p_t)|] < \Delta/2 + M_{t,\Delta}\theta_t^{c,0}, a.s.$$

If a weaker boundedness assumption (iv)′′: Posterior state space is compact (not uniformly compact) holds, then we have

$$\lim_{N\to\infty} \Xi_{pf}[|K(\pi_t^{c,0,N} : p_t^{M_{t,\Delta}}) - K(\pi_t^c : p_t)|] < M_{t,\Delta}\theta_t^{c,0}, a.s.$$

[8] Note p_t^M is not a pdf.

[9] Assumptions (ii) & (iii) imply that $\rho_k^0 < \infty, a.s.$ (Remark 5.6 of [3]).

V. SLOW AND DRASTIC CHANGES: ELL AND OL

A. The OL Statistic

As discussed earlier, the drastic change detection problem is well studied in literature and algorithms like CUSUM [6] which are based on the likelihood ratio of observations can be used wherever it can be evaluated. When change parameters are unknown, the likelihood ratio can be replaced by negative log likelihood of current observation given past observations, which we call **observation likelihood (OL)**, $OL = -\log P(Y_k|Y_{1:k-1}, H_0)$. A change is declared if OL exceeds a threshold. OL is evaluated using a PF for the given PONLD model (Section I-B) as $OL_k^N = -\log(Q_k^0 \pi_{k-1}^N, \psi_k)$.

Now, if the change is drastic, the likelihood of observations under the normal (unchanged) model will reduce (OL which is its negative log will increase) or equivalently the particle filter, which is optimal for the normal system, will lose track. Thus OL can be used to detect this change. But due to asymptotic stability [3], the particle filter is able to track slow changes and hence these are missed by OL. We show below using the theorems from the previous section, that such slow changes are picked up by ELL, which in fact can be estimated correctly for the changed system (using a PF optimal for the normal system) only because of asymptotic stability.

B. Comparing ELL and OL Performance

Consider the finite time situation (fix $t \le T$ for some large T) and apply theorem 5. Set $M = \max_{1\le t\le T} M_{t,\Delta}$, $N = \max_{1\le t\le T} N_{t,M_{t,\Delta},\Delta}$. Then we have

$$\Xi_{pf}[|K_t^0 - K_t^{0,M,N}|] < \Delta/2 + \frac{M\beta_t^0}{\sqrt{N}}$$
$$\Xi_{pf}[|K_t^c - K_t^{c,0,M,N}|] < \Delta/2 + \frac{M\beta_t^{c,0}}{\sqrt{N}} + M\theta_t^{c,0} \qquad (12)$$

where $\beta_t^0 = \beta_t(\rho_k^0, \epsilon_k^0, 0 \le k \le t)$, $\theta_t^{c,0} = \theta_t(\delta_k^{c,0}, \epsilon_k^c, t_c \le k \le t)$, and $\beta_t^{c,0} = \beta_t(\rho_k^0, \epsilon_k^0, 0 \le k \le t_c, \rho_k^{c,0}, \epsilon_k^{c,0}, t_c \le k \le t)$ and θ_t, β_t defined in (8), (10) respectively. First consider the PF error. Although theoretically, it can be made to decrease to zero, with $N \to \infty$, in practice it is the most dominant source of error. For normal system observations, it is the only source of error and for changed system observations, this is because it is not possible to fix a value of N to ensure a certain maximum error ($\beta_t^{c,0}$ is not known). We can only choose N large enough to have the error small for normal observations (H_0). But when tracking observations coming from H_c using model H_0, a much larger N is required. Now the PF error coefficient $\beta_t^{c,0}$ depends on past values of $\epsilon_k^{c,0}$ and $\rho_k^{c,0}$. Using Remark 5.10 of [3], we have the following upper and lower bounds on ρ_k which can be expressed in terms of $OL_k^{c,0}$:

$$\frac{\sup_{x\in E_{x,y}} \psi_k^c(x)}{(Q_k^0\pi_{k-1}^{c,0}, \psi_k^c)} \le \rho_k^{c,0} \le \frac{\sup_{x\in E_{x,y}} \psi_k^c(x)}{(\epsilon_k^{c,0})^2 (Q_k^0\pi_{k-1}^{c,0}, \psi_k^c)}$$
$$\implies \frac{\sup_{x\in E_{x,y}} \psi_k^c(x)}{e^{-OL_k^{c,0}}} \le \rho_k^{c,0} \le \frac{\sup_{x\in E_{x,y}} \psi_k^c(x)}{(\epsilon_k^{c,0})^2 e^{-OL_k^{c,0}}} \qquad (13)$$

Now consider the model error, $\theta_t^{c,0}$. It depends on past values of $\delta_k^{c,0}$ and ϵ_k^c. ϵ_k^c is a constant which depends only on the mixing properties of R_k^c. Using a slightly modified version of theorem 3 of our recent work [18], we can bound $\delta_k^{c,0}$ in terms of $OL_k^{c,0}$:

$$\delta_k^{c,0} \le \frac{2D_{Q,k}}{e^{-OL_k^{c,0}}} \qquad (14)$$

where $D_{Q,k} = \sup_x \int_E \psi_{t,Y_t^c}(x')|Q_t^c(x,x') - Q_t^0(x,x')|dx'$ is defined in [18] as a metric for the rate of change (change magnitude per time step). Now we have the following observations:

- For a small change magnitude per time step (small $D_{Q,k}$), $OL^{c,0}$ will not be significantly larger than OL^0 and hence *OL may not be able to detect the change or may take long to detect it*. But by (13) and (14), this also implies that the upper bounds on ρ_k and δ_k are smaller or that the PF and model error in approximating ELL are small. Thus, in this case, *ELL will be able to detect the change*. Assuming negligible errors, with probability greater than $(1 - 0.11) = 0.89$, ELL detects the change at or before time t for which $\gamma_t > 3\sqrt{VK_t^c}$ (from Section III-B).
- From (13), the upper bound on ρ_k is inversely proportional to $(\epsilon_k^{c,0})^2$ and by theorem 2, β_t is also inversely proportional to past values of $(\epsilon_k^{c,0})^2$. Thus *PF error upper bound is inversely proportional to* $(\epsilon_k^{c,0})^4$. Now, *the magnitude of* $\epsilon_k^{c,0}$ *depends inversely on the total magnitude of change*. For example, in Example 1, assume that $\psi_k(x)$ has finite support, i.e. $\psi_k(x) = 0$, $\forall |Y_k - h(x)| > B$. This can be achieved for example if the observation noise is a truncated zero mean Gaussian truncated at $\pm B$. This assumption makes the kernels $R_k^0, R_k^c, R_k^{c,0}$ mixing (Example 3.10 of [3]). Also let $h(x) = x^2$, then (using Example 3.10 of [3]) $\epsilon_k^{c,0} = e^{-2(max(Y_{k-1}^c, Y_k^c)+B)}$. Now $E[Y_k^c] = a_k^2 + \sigma_k^2$, so that as the change magnitude, a_k, increases, the random variable $\epsilon_k^{c,0}$ decreases (stochastically) and consequently the PF error increases (stochastically).

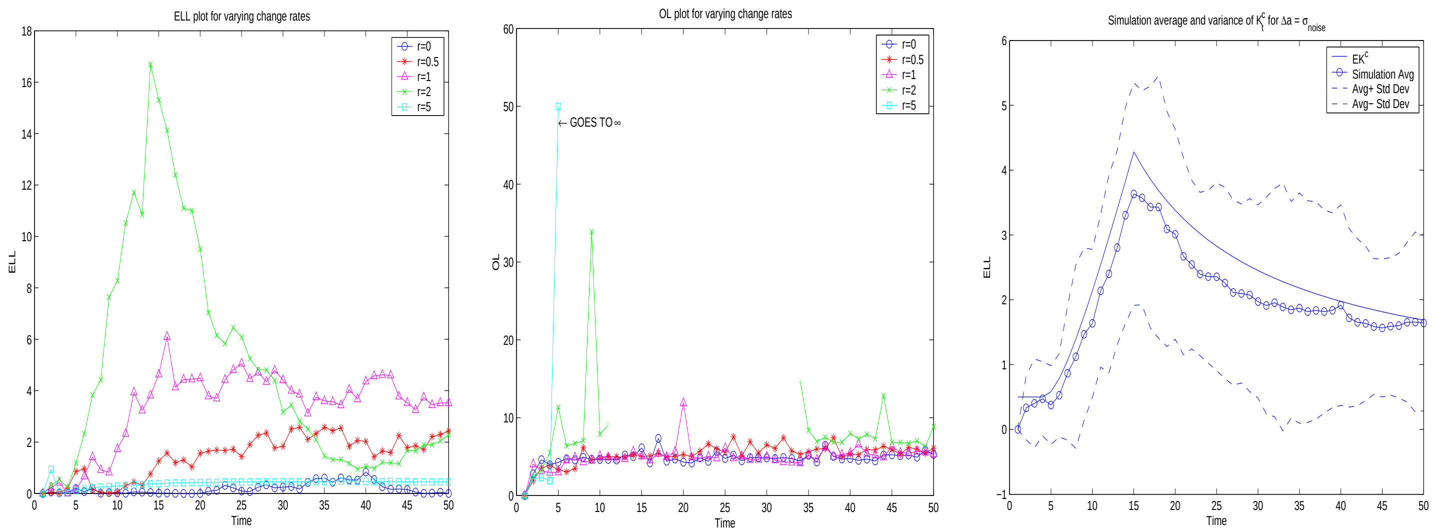
Usually when using a PF, one of the following happens: Either the change is slow enough so that the PF does not "completely lose track" until γ_t is large enough for the change to get detected. Or, if the change is not slow enough, the PF "completely loses track" but in that case, OL will detect the change. Thus, we *propose to use a combination of ELL and OL for change detection in PONLD systems (when the rate of change can be slow or fast and change parameters are unknown)*. A change should be declared when either exceeds its respective threshold.

VI. SIMULATION RESULTS

We simulated Example 1 with $\psi(x)$ having compact support (truncated Gaussian) and taking $h_t(x) = x^2$. We tested for increasing magnitudes of Δa. We tested for $\Delta a = r\sigma_{noise}$ with $r = 0$ (no change) and $r = 0.5, 1, 2, 5$. We show in Figure 1, plots for detecting the changes using ELL and OL. As can be from these graphs all changes are detected by either OL or ELL. The "slow" changes ($r = 0.5, 1$) are missed by OL but detected by ELL[10]. The "faster" change ($r = 2$) gets detected by both although ELL detects it faster. The "drastic" ($r = 5$) change gets missed by ELL but OL detects it immediately. Also note that when OL takes the value infinity (due to overflow), ELL starts to fail. The $r = 5$ (cyan-square) ELL plot in Figure 1(a) almost coincides with that of $r = 0$ (normal system). This is because when PF loses track, the posterior starts following the normal system model, i.e. $R_t^c \approx Q_t^0$.

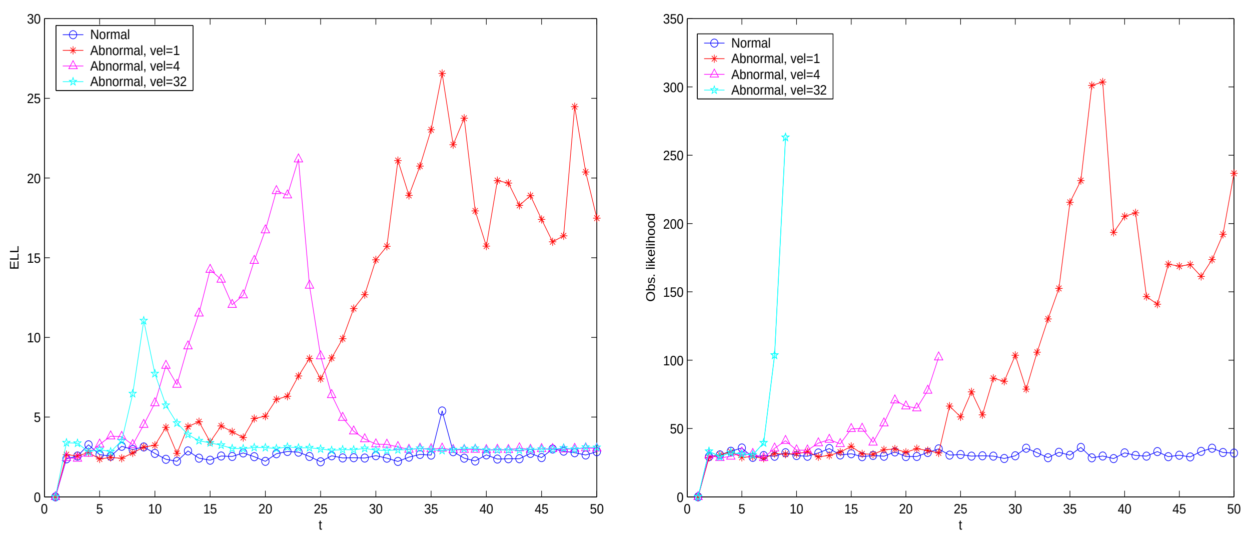
Now we also show application of our change detection strategy to a computer vision problem of abnormal activity detection [1], in which we modeled the normal activity using a PONLD system. In [1], we proposed a (stochastic) shape dynamical model for

[10] The change with $r = 0.5$ and duration only 10 time units ($t_c = 5, t_f = 15$) is too small for ELL to detect and in many of the realizations that we simulated, this change was not detected at all.



(a)ELL plot (b)OL plot (c)Analytical & Simulation average and spread

Fig. 1. Simulated example: In (a) and (b), we show ELL and OL (negative log of observation likelihood) plots for the no change case (blue -o), and for changes with $\Delta a = r\sigma_{noise}$ for r=0.5 (red-*), r=1 (magenta -△), r=2 (green -x) and r=5 (cyan -$square$). In all cases change was introduced at time $t_c = 5$ and lasted till $t_f = 15$. For the case $r = 5$ (drastic change), the OL plot goes to infinity after $t = 5$ (computer overflow) and hence the change is detected immediately using OL while ELL completely fails for it. The $r = 2$ ("faster change") gets detected at $t = 9$ using OL but ELL detects it at $t = 6$ itself. The slower changes $r = 0.5, 1$ get detected by ELL but are missed by OL. In (c), we plot $0.5a_t^2/\sigma_t^2$, its simulation average calculated using 20 realizations of the observation sequence and its spread (average plus and minus the standard deviation) for a change with $\Delta a = \sigma_{noise}$.



(a) ELL (b) OL

Fig. 2. We show in (a) and (b), plots of ELL and OL for normal activity and increasing walk away velocities (abnormal behavior) as a function of time. Abnormality is introduced at $t = 5$. The $vel = 32$ ("drastic change") plot of OL goes to infinity (overflow) at $t = 5$ and hence abnormality gets detected immediately, and for $vel = 4$ ("faster change"), the OL plot goes to infinity at $t = 24$. For all changes except $vel = 32$, ELL detects faster.

modeling the changing configuration of a group of moving objects. The observations of the object locations obtained using an automated motion detection algorithm are noisy, making the system partially observed. In the specific application we considered, we modeled the "normal activity" of a group of passengers deplaning and moving towards the terminal in an airport (See [1] for images of the normal and abnormal activity). The shape and motion at time t constituted the state vector, X_t. *Abnormality detection was formulated as a change detection problem with change parameters unknown.* We studied the problem of detecting the change in the shape due to one person walking away from his normal path in some other direction. The speed at which the person walked away decided the rate of change. We show in Figure 2, the plots of ELL and OL to detect the abnormality for increasing rates of change (walk-away velocities). As before, velocity=1 was a slow change which got detected by ELL much faster than OL, while for velocity=32, ELL failed and OL detected immediately.

VII. Conclusions and Future Work

We have proposed a change detection statistic, ELL, for slow change detection in PONLD systems tracked using particle filters and have studied errors in its approximation (modeling error in tracking changed observations using original system transition kernel and PF approximation error). We have proved in Section IV, the asymptotic convergence of the errors to zero as $M, t, N \to \infty$. Slow changes are missed by tracking error or observation likelihood (OL) because the PF is able to track the slow change due to asymptotic stability. But on the other hand, we have shown that ELL is able to detect slow changes because of asymptotic stability. We have discussed in Section V, this complementary behavior of ELL and OL for change detection, using the results from Section IV. Simulation results on a one dimensional problem and a real abnormal activity detection application have been presented to support our theoretical claims.

As part of future work, we hope to prove convergence of $M_{t,\Delta}\theta_t$ to zero as $t \to \infty$, using only the assumptions of theorem 5. We also intend to study practical examples of non-linear systems which satisfy the assumptions required for applying theorems 3, 4 and 5. Also, in the analysis in this paper, the error in the approximation of K_t^c by $K_t^{c,0,N}$ is much larger than that of K_t^0 because we are using a filter which is optimal for the original system. But if one were to make the transition kernel used in the filter less specific, for example in the case of Example 1, use σ^2_{noise} larger than the true variance of Q_t^0, it will make $R_t^{c,0}$ more mixing and thus reduce the approximation error of K_t^c without significantly increasing error in estimating K_t^0. This has been observed experimentally. We have analyzed this problem in a recent work [18] where we show that the model and PF errors in estimating any function of the state are upper bounded by increasing functions of the system model error per time step (here rate of change). Finally, we also intend to study the performance of a CUSUM [6] like algorithm applied to ELL (use ELL of a subsequence of past states).

References

[1] N. Vaswani, A. RoyChowdhury, and R. Chellappa, "Activity recognition using the dynamics of the configuration of interacting objects," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[2] N.J. Gordon, D.J. Salmond, and A.F.M. Smith, "Novel approach to nonlinear/nongaussian bayesian state estimation," *IEE Proceedings-F (Radar and Signal Processing)*, pp. 140(2):107–113, 1993.

[3] LeGland F. and Oudjane N., "Stability and Uniform Approximation of Nonlinear Filters using the Hilbert Metric, and Application to Particle Filters," *Technical report, RR-4215, INRIA*, 2002.

[4] Rudolf Kulhavy, "A geometric approach to statistical estimation," in *IEEE Conference on Decision and Control (CDC)*, Dec. 1995.

[5] D.F. Kerridge, "Inaccuracy and inference," *J. Royal Statist. Society, Ser. B*, vol. 23 1961.

[6] M. Basseville and I Nikiforov, *Detection of Abrupt Changes: Theory and Application*, Prentice Hall, 1993.

[7] B. Azimi-Sadjadi and P.S. Krishnaprasad, "Change detection for nonlinear systems: A particle filtering approach," in *American Control Conference*, 2002.

[8] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*, Academic Press, 1988.

[9] Shaohua Zhou and Rama Chellappa, "Probabilistic human recognition from video," in *European Conference on Computer Vision*, 2002, pp. 681–697.

[10] D. Ocone and E. Pardoux, "Asymptotic stability of the optimal filter with respect to its initial condition," *SIAM Journal of Control and Optimization*, pp. 226–243, 1996.

[11] Rami Atar and Ofer Zeitouni, "Lyapunov Exponents for Finite State Nonlinear Filtering," *SIAM Journal on Control and Optimization*, vol. 35, no. 1, pp. 36–55, 1997.

[12] F. LeGland and L. Mevel, "Exponential forgetting and geometric ergodicity in hidden markov models," *Mathematics of Control, Signals and Systems*, pp. 63–93, 2000.

[13] R. Atar, "Exponential stability for nonlinear filtering of diffusion processes in a noncompact domain," *Ann. Probab.*, pp. 1552–1574, 1998.

[14] A. Budhiraja and D. Ocone, "Exponential stability of discrete time filters for bounded observation noise," *System and Control Letters*, pp. 185–193, 1997.

[15] P. DelMoral, "Non-linear filtering: Interacting particle solution," *Markov Processes and Related Fields*, pp. 555–580, 1996.

[16] G. Casella and R. Berger, *Statistical Inference*, Duxbury Thomson Learning, second edition, 2002.

[17] H.L. Royden, *Real Analysis*, Prentice Hall, 1995.

[18] N. Vaswani, "Bound on errors in particle filtering with incorrect model assumptions and its implication for change detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.